A STUDY ON PROBABILISTIC AND COMPUTATIONAL APPROACHES TO RISK MODELING, ANALYSIS AND FORECASTING

By

NOH-JIN PARK

Bachelor of Arts in Applied Statistics
Yonsei University
Seoul, Korea
1982

Master of Science in Statistics
Seoul National University
Seoul, Korea
1984

Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
DOCTOR OF PHILOSOPHY
July, 2009

A STUDY ON PROBABILISTIC AND COMPUTATIONAL APPROACHES TO RISK MODELING, ANALYSIS AND FORECASTING

Dissertation Approved:

Dr. K. M. George

Dissertation Adviser

Dr. Venkatesh Sarangan

Dr. Xiaolin (Andy) Li

Dr. Jaebeom Kim

Dr. A. Gordon Emslie

Dean of the Graduate College

ACKNOWLEDGMENTS

I would like to thank Professor K. M. George, my graduate advisor and mentor, for letting me freely and independently work on my research problems. Professor K.M. George's patience and encouragements always helped stay alert. With his valuable and family-like supervision on my doctoral study, I could establish and enhance my knowledge in computationally-based statistical risk analysis and forecast. I also feel very fortunate that I had the opportunity to learn from Dr. Venkatesh Sarangan, Dr. Xiaolin (Andy) Li, Dr. Jaebeom Kim who provided me valuable suggestions on my doctoral work. I also would like to thank my colleagues and friends, especially Professor JongWoo Jeon and HaeJin Shin, for their support throughout my study. I would like to thank Dr. Nohpill Park for showing me the way to the field of computer science. I would like to share this accomplishment and joy with My daughter, YoonJung.

Lastly, I devote this dissertation to My Parents, My Mother, Inbok Lee, who is always praying for my success in study and My Late Father, Professor HongNai Park, who must be the happiest one than anyone.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Risk analysis has been extensively studied to address and resolve the issues related to risk assessment, risk characterization, risk communication, risk management, and risk-related policies [26, 27]. Risk analysis can be formally defined by a science that deals with probabilistic and statistical risk evaluation. Probabilistic or statistical risk assessment can be defined as an analysis methodology employed in science and engineering where risk can be evaluated by its probabilistic behaviors [26]. Therefore, risk analysis is a methodology to provide a way to assess, manage and forecast risks in probabilistic processes in an integrative manner. Risk analysis is the topic of this study with focus on data with turbulent trends.

The Ad-Hoc Risk Management System (ARMS) [1] has been developed as a base tool for statistical modeling and analysis of data, and forecast of risky events or events with a turbulent nature as a primary focus. The ARMS is based on the Multiple Regression with a Scaled Dummy Variable (MRSDV) Model [1], and primarily aims at such events as critical network security breaches, terror events and financial market turbulence, to mention a few. MRSDV provides a capability to identify a spiky pattern in historical data and is extensible to make a forecast on spiky events in the

future based on the historical trend. However, MRSDV limits itself due to a few drawbacks such as no efficient method for parameter estimation, no adequate test scheme for evaluation of model goodness-of-fit, and no solid criterion between risky and non-risky events.

In order to address and resolve the drawbacks and issues in MRSDV, the Multiple Regression with a Dependent Dummy Variable (MRDDV) Model [2] has been developed by employing the following three techniques: a dependent dummy variable technique, the Ordinary Least Square (OLS) method for efficient parameter estimation, and the coefficient of determination $\left(R^2\right)$ as a test scheme for evaluation of model goodness-of-fit. MRDDV employs the multiple regression method which manipulates the quantitative independent variable and the qualitative independent variable, along with a dependent dummy variable in order to establish a solid definition and criterion for a risk state. Based on the risk state, the MRDDV method is capable of detecting likelihood of risky events. The MRDDV is extended to be able to compute the likelihood of such risky events by a probability with respect to the distribution of the estimated values of the dependent dummy variable in the regression model. The MRDDV assigns a dummy binary value (e.g., 0 or 1) to its dependent dummy variable as a criterion function, such that 1 is assigned if the value of an independent variable, either the quantitative or qualitative variable, intersects a threshold value set by the user. However, this limits its extensibility to a probabilistic exploration of the risk process by a risky event with poor model goodness-of-fit. In order to improve the level of model goodness-of-fit for risky events without any loss of generality, and also to resolve the detectability or observability of risky events hit

in a distributed manner observed far away from the threshold value of interest in MRDDV, in this study, it is proposed that the values of the dummy dependent variable be adjusted to the values determined by a new criterion function in which the threshold or criterion value is defined by degree of difference between adjacent data values instead of a single threshold value every data value is compared against. This way, the risky or turbulent events can be detected and observed as an event that reveals an abrupt change in the value of the process at any period over lifetime of the process. Therefore, an event that intersects a threshold value yet without revealing an abrupt change in value will not be identified as a risky event by using ordinary MRDDV while it will be identified as a risky event by using the new proposed approach.

The new MRDDV with the adjustment in the criterion function for determining the dependent variable value is referred to as Multiple Regression with Adjusted Dummy Variable (MRADV). Both MRDDV and MRADV can be extended to be able to measure the probability for a risky event to occur; MRDDV provides an insight into the likelihood for such event to go beyond a certain threshold value while MRADV traces distribution of abrupt changes in values of the random variable of interest.

The probabilistic MRADV, referred to as P_{MRADV}, can be implemented in two different ways such as Observation-Driven Method (ODM), in which the probability is measured based on the estimated values of the dependent adjusted variable; and Input-Driven Method (IDM), in which the probability of the risk of concern is traced based on both the estimated independent quantitative and qualitative random

variables. The effectiveness and computational efficiency of P_{MRADV} is demonstrated by comparing it against Copula based method to compute joint probability.

The novelty of the proposed P_{MRADV} is that by using traditional methods such as Copula-based probabilistic modeling and analysis method, the probability density functions for the quantitative independent variable and qualitative independent variable can only find a joint probability without being able to take into account their joint-impact on a common figure of merit such as the dependent dummy or adjusted variable; while P_{MRADV} or P_{MRDDV} can effectively take into account and evaluate the joint-impact of the two independent random variables with respect to a threshold value which is determined by the user by using a criterion function. The criterion function can be set simply to be based on the value of the quantitative independent random variable as is in P_{MRDDV}; based on the changing-rate in the values of the variable as is in P_{MRADV}; or based on a composite function to involve both the quantitative and qualitative independent random variables. Hence, a method to evaluate the joint-impact is made possible by the multiple regression process in P_{MRDDV} or P_{MRADV} during which the individual impact from the qualitative independent random variable is made visible to the dependent random variable either implicitly (if the qualitative independent random variable did not participate in the criterion function) or explicitly (if the qualitative independent random variable is included in the criterion function as an explicit variable). Ultimately it will be made possible to predict the likelihood for an event (or a risky event) of interest (particularly, such as an event with an abrupt or spiky turbulence in its value) to occur with reference to

other events that are independent contributors to the occurrence of the risky turbulence in the primary event, which otherwise may stay hidden from other traditional prediction methods.

Lastly, in this dissertation, a method is developed that can facilitate the extension of the Multiple Regression with Dependent Dummy Variable (MRDDV) Model to provide a way of estimating the likelihood of any event of concern or interest by probability. MRDDV Model employs a dependent dummy variable as an observation in its regression model with respect to the quantitative independent variable and the qualitative independent variables as primary inputs for estimation. The purpose of the dependent dummy variable in MRDDV is to provide an effective way of representing the quantitative measure of the status of the event of concern with respect to a certain criterion function, such as a binary measure (e.g., *0* or *1*) or forward differences of dependent variable values. Therefore, MRDDV can facilitate the process of identifying the quantitative relationship among the random variables in the model by using the regression-based estimation. However, MRDDV lacks the ability to readily provide information on how likely an event of concern is to occur, which could be best manipulated by employing probability-based estimation. In this context, a method, namely Logit Transformation, is employed to facilitate the probabilistic manipulation of MRDDV. By using Logit Transformation method, the estimated dependent dummy variable can be transformed from a non-probabilistic domain (e.g., the estimated value could be out of valid probabilistic range, *0* to *1*) into a probabilistic one so the expected value of the dependent dummy variable can be evaluated as a probabilistic measure. Applying this Logit Transformation to original

MRDDV model, we specify three procedures as follows: i) a variable $P = \Pr(y_i = 1)$ as a probability of dependent dummy variable (y_i) takes 1, ii) set $\log\left(P/(1-P)\right)$ equal to right hand side of original MRDDV model, and iii) arrange the formula in terms of P. Then we can get the Logit Transformation-based MRDDV model. The Logit Transformation-based MRDDV [4] is validated by applying to historical financial data.

The remainder of this dissertation is organized as follows: In chapter II, literature related to risk analysis is reviewed, previous works are also introduced. The basic principles, methodologies and details of the proposed model are described in chapter III. The performance of the proposed method is evaluated in chapter IV. In the final chapter, remarks and conclusion are presented.

CHAPTER II

LITERATURE REVIEW

ARMS [1] is introduced and employed as a base tool for risk analysis and management as it can provide an integrated process of risk management and assessment in the fields in which an integrated risk analysis and forecast is exigently required to tolerate or avoid any consequential undesirable impact. Thus, the objective of ARMS is to estimate/predict possible outcomes that are inestimable and unpredictable directly from the observations in a data space. In order to estimate and predict the situations in advance, not only historical data but also the intelligence of experts in the field is essential. The basic process of risk analysis is to construct a model to estimate or forecast events in a given environment. Traditional estimation theory such as regression-based estimation method and reliability distribution-based estimation method can perform estimation in an ad hoc manner, however they cannot perform properly and adequately under real risky events or situations because they calculate the estimator or make a forecast based only on historical data which is a quantitative variable.

In [1], a statistical model for predicting abnormal spikes in a data space has been proposed. Based on the proposed statistical model in [1], an Ad-hoc Risk

Management System (ARMS) was developed and the proposed ARMS facilitates statistically and computationally the process to develop an alert system against unexpected and disastrous events such as terror attack and abrupt stock market fluctuation, to mention a few. One of the most important issues in ARMS is the timing when to watch and warn against the risk.

The statistical model proposed in [1] is referred to as the Multiple Regression with Scaled Dummy Variable (MRSDV), and it addresses and resolves how to model and forecast the time of risky events. MRSDV evaluates a set of data in which a spike-pattern is introduced as a sign of a risky event. In general, traditional estimation theories [5, 6, 7] can not readily manipulate the estimation adaptively in a changing environment. Estimation models may be used for an ad-hoc risk management by using multiple regression and the Weibull distribution-based approach [3] as well as various time series models [12], however, they can not perform properly and adequately under real risky situations because they calculate its estimator or forecasting value based only on the historical data which is a quantitative variable. MRSDV model manipulates the estimation with respect to both the quantitative independent variable and qualitative independent variable, in which the qualitative variable is surveyed by using a method such as k-point scaling to introduce and generate a dummy variable [1, 8, 9, 11]. MRSDV is used as the basic statistical method of the proposed model [5, 6, 7]. In [1], MRSDV was compared with the conventional methods that forecast an event based only on historical data, to show the effectiveness of the proposed predicting method using scaled dummy variables, in terms of accuracy and precision.

MRSDV in [1] has left some unanswered questions. As can be seen in [1], MRSDV is able to capture a spike pattern of two homogeneous data well if there exists one. However, MRSDV model can not effectively provide a way to specify or set a criterion of risk/ non-risk because it is not readily possible to derive a probabilistic measure of the likelihood of an event of concern or interest with respect to a given criterion or threshold value based on the regression model estimated in MRSDV, and does not support an efficient parameter estimation because the OLS (Ordinary Least Square) method can not be applied to nonlinear regression model such as MRSDV [5, 7]. In order to overcome these limitations of [1], a dependent dummy variable technique has been proposed in [2], where the MRDDV (Multiple Regression with Dependent Dummy Variable) Model was developed. MRDDV Model also employs the multiple regression method and manipulates the estimation with respect to both the quantitative independent variable and qualitative independent variable as in MRSDV. However in MRDDV, instead of introducing a scaled dummy variable and qualitative variable, a technique is proposed to introduce a dummy variable which is dependent of the quantitative and qualitative variables in order to set a criterion for evaluating a risk state. Furthermore, MRDDV facilitates its parameter estimation by employing the Ordinary Least Square (OLS) method. Also, in [2], a technique for model goodness-of-fit test was presented.

MRDDV model employs the multiple regression method which runs between the quantitative independent variable and the qualitative independent variable, with dependent dummy variable technique to set the criterion of risk state. It improved the criterion for risk state, parameter estimation method and test scheme for model

goodness-of-fit than MRSDV Model, and also attempts to represent risk level based on probability. Traditionally, probabilistic risk modelling and analysis has been conducted by defining a joint distribution in terms of marginal and conditional distributions for the model's random variables such as in copula [10, 13-18]. Copula is a technique to construct the joint distribution function between independent variables. In this study, a probabilistic approach of estimated response of Adjusted MRDDV Model is proposed in two different views such as Observation Driven Method (ODM) and Input Driven Method (IDM). Risk management is a human activity which integrates recognition of risk, risk assessment and developing strategies to manage it.

The basic aspect of risk analysis is constructing the model to estimate or forecast events in related environment. There are several facets of risk analysis in the literature. In financial market risk measurement, there is the technique of context modelling based on Value-at-Risk (VaR). It suffers from major drawbacks. The log-normal modelling of the returns does not take into account the observed fat tail distribution and the non-stationality of the financial instruments sever the efficiency of the VaR predictions. The technique of context modelling is applied to estimate the VaR by conditioning the probability density function on the present context [21]. Based on Markowitz's portfolio selection model that includes proportional transaction costs in the presence of initial holdings for the investor, there is a risk minimization model that is a portfolio optimisation module (PORTOPT) that computes the income produced by the portfolio during the horizon period and computes market value at the end of the horizon [22]. Also, there is a transaction method for risk mapping which allows one to compute confidence

intervals for estimates, without any assumption on data distribution, except that inputs should be independently and identically distributed. RRCM (Ridged Regression Confidence Machine) gave a good numerical performance only with ‘iidness (identical and independent)’ of data coordinates [23]. In risk analysis, finally, the use of probabilistic techniques to address variability and uncertainty in risks is introduced, focused on how better characterization of the variability and uncertainty in the risk assessment lead to better risk communication [24].

In this study, ARMS is adopted as it is the mixed feature of risk management and risk assessment, in the fields too sensitive to tolerate risks that might result in severe disaster. There are several example scenarios to which a risk management system could be applied to forecast and avoid disaster. One such example is the field of terrorism. The basic aspect of risk analysis is constructing the model to estimate or forecast events in related environment.

CHAPTER III

METHODOLOGY

This dissertation presents a statistically-based yet probabilistically-concluded approach to modeling and evaluation of likelihood for events of interest to occur with a focus on risky events. The risky events of interest in this study are the ones with a turbulent nature in the distribution of values of data, which can be commonly found in the events in the fields such as financial market, homeland security, or safety/mission critical systems, to mention a few. In such events, it is critical to make a timely, practical and accurate forecast for the likelihood of the events of interest to occur. There have been a couple of methods developed, i.e., Multiple Regression with a Scaled Dependent Variable (MRSDV) [1] and Multiple Regression with a Dependent Dummy Variable (MRDDV) [2]. They are both multiple regression method-based, in which a quantitative and a qualitative variables are employed to represent inputs along with an output variable to represent the consequence of the inputs on the observation through the regression process. What distinguishes MRSDV and MRDDV is the function that determines the values for the dependent variable in their models, referred to as criterion function. Using a criterion function, the dependent variable in MRSDV may result in an infinite positive range, while that in MRDDV

may result in a binary output such as 0 or 1, respectively. The new approach proposed in this study is based on a new criterion function applied into the multiple regression process, referred to as Multiple Regression with an Adjusted Dependent Variable (MRADV). In MRADV, the adjustment on the dependent variable is made by determining the value of the dependent variable based on the absolute values of the forward difference of adjacent data values in the quantitative random variable in order to improve the model goodness-of-fit. Furthermore, based on the multiple regression model in the proposed MRADV, a probabilistic-based model, referred to as P_{MRADV}, is proposed in order to derive the probability of risk (or an event of interest) to occur without relying on traditional way of assuming or establishing probability density functions of the random variables in the model, thereby guiding the users to a more practical and realistic evaluation of the likelihood of an event of interest to occur. Extensive simulations and thorough interpretation of the results is conducted and presented to demonstrate the validity of the proposed approach by comparing it against the traditional copula-based approach in its computing joint probability.

Lastly, in this dissertation, a method is presented, that can facilitate the extension of the Multiple Regression with Dependent Dummy Variable (MRDDV) Model to provide a way of estimating the likelihood of any event of concern or interest by probability. MRDDV Model employs a dependent dummy variable as an observation in its regression model with respect to the quantitative independent variable and the qualitative independent variable as primary inputs for estimation. The purpose of the dependent dummy variable in MRDDV is to provide an effective way of representing the quantitative judgment of the status of the event of concern with respect to a

certain criterion function, such as a binary judgment (e.g., *0* or *1*) or forward differences of independent variable values, to mention a few. MRDDV can facilitate the process of identifying the quantitative relationship among the random variables in the model by using the regression-based estimation. However, MRDDV lacks the ability to readily provide information on how likely an event of concern is to occur, which could be best manipulated by employing probability-based estimation. In this context, a method, namely Logit Transformation, is employed to facilitate the probabilistic manipulation of MRDDV. By using Logit Transformation method, the estimated dependent dummy variable can be transformed from a non-probabilistic domain (e.g., the estimated value could be in the range beyond *0* or *1*) into a probabilistic one so the expected value of the dependent dummy variable can be evaluated as a probabilistic measure. The various models are described in the following subsections.

3.1 MRSDV Model

In this section, we describe the statistical model, referred to as Multiple Regression with Scaled Dummy Variable (MRSDV) is detail. A numerical simulation result is provided to verify the practicality and effectiveness of the proposed method in Chapter IV. The theoretical foundation for MRSDV model using k-point scaling as its dummy variable can be expressed as follows.

$$y_{ij} = \alpha_0 + \alpha_1 \sum_{j=1}^{n} x_{1ij} + \beta_1 \sum_{j=1}^{n} x_{2ij} + \gamma_1 \sum_{j=1}^{n} x_{1ij} \sum_{j=1}^{n} x_{2ij} + \varepsilon_{ij} \qquad (3.1)$$

where y_{ij} : risk function (response)

x_{1ij} : quantitative independent variable (*jth* observation in *ith* time period)

x_{2ij} : qualitative independent variable (*jth* mean from *ith* experts group)

α_0 : intercept of the model

α_1 : coefficient of quantitative variable (parameter of *ith* time perioid)

β_1 : coefficient of qualitative variable (parameter of *ith* experts group)

γ_1 : coefficient of product of quantitative and qualitative variable (parameter of ith interaction between historical data and dummy data)

i : time lag

ε_{ij} : error term

3.1.1. Stepwise selection of quantitative variable

To conduct MRSDV model, firstly, we have to select the variable to fit the response adequately. Hence, we have to find out the variable highly correlated with the response. For example, the expense-amount of terror group, the number of their phone calls or reservation of air or ground transportation have to be found out as an independent variable highly correlated with the terror occurrence. At this point we need a reasonable variable selection method. For MRSDV model we adopt stepwise regression that is an automatic variable selection procedure based on F-testing method [5]. There are three major approaches such as i) forward selection involves starting with no variables in the model, trying out the variables one by one and including them if they

are statistically significant. ii) backward elimination involves starting with all possible variables and testing them one by one, and deleting any of them if they are not statistically significant. iii) stepwise selection tests at each stage for variables to be included or excluded based on forward selection and backward elimination.

3.1.2. Dummy variable

To include the qualitative variable which is observed or collected by nominal scale or k-point scaling in MRSDV model, a dummy variable is introduced.

For example, let y be the sales amount of beverage in certain soccer game, x_1 be the number of tickets sold out and x_2 be the quality of the game, then the regression model can be expressed as follows;

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

where $x_2 = 0$ if it is not the game of top rankers,

$=1$ if it is the game of top rankers

Here we refer the variable x_2 as the dummy variable and its response function is as follows.

$$E(y) = \beta_0 + \beta_1 x_1 \qquad \text{if } x_2 = 0$$

$$= (\beta_0 + \beta_1) + \beta_1 x_1 \qquad \text{if } x_2 = 1$$

3.1.3. Standardization of variable

In case of MRSDV as well as multiple regression, we have to standardize the observation because they may have different bases. As a result of this, in general, the normalized variable has variance 1 [1, 2].

Assume, for example, multiple regression with two independent variables which have different bases as follows;

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

where y: sales amount (value)

x_1: advertise amount (value)

x_2: store size (square feet)

Here we have to standardize y, x_1, x_2 respectively such that;

$$y^* = \frac{y - \bar{y}}{s_y}, \quad x_1^* = \frac{x_1 - \bar{x}_1}{s_1}, \quad x_2^* = \frac{x_2 - \bar{x}_2}{s_2}$$

where s: standard deviation of y

$\bar{y}$: mean of y

s_1: standard deviation of x_1

$\bar{x}_1$: mean of x_1

s_2 : standard deviation of x_2

$\bar{x}_2$: mean of x_2

Then, new regression model can be derived as follows, with standardization with same base.

$$y^* = \beta_0{}^* + \beta_1{}^* x_1{}^* + \beta_2{}^* x_2{}^* + \varepsilon^*$$

This is the MRSDV model in two variables.

3.1.4. Risk management in MRSDV model

The objective of ad hoc risk management is to predict an outcome that is significantly different from the observations in a data space. To predict the situations in advance, not only historical data but also the knowledge and intelligence of experts in that field are prerequisite. Now, we introduce a technique to insert these information to MRSDV model.

3.1.5. SWOT analysis of qualitative variables

Making a decisive prediction is always difficult because there is a significant gap between yes and no. Here we propose a simple tool to make the prediction easier in complicated circumstances, which is referred to as SWOT (Strength, Weakness, Opportunity and Threat) analysis [8]. An adjusted SWOT analysis is used in the

proposed MRSDV model as shown in Figure 3.1 switching the terminology 'Opportunity' to 'no threat'. Figure 3.1 shows the implementation of the proposed adjusted SWOT and explain it by using the SWOT map.

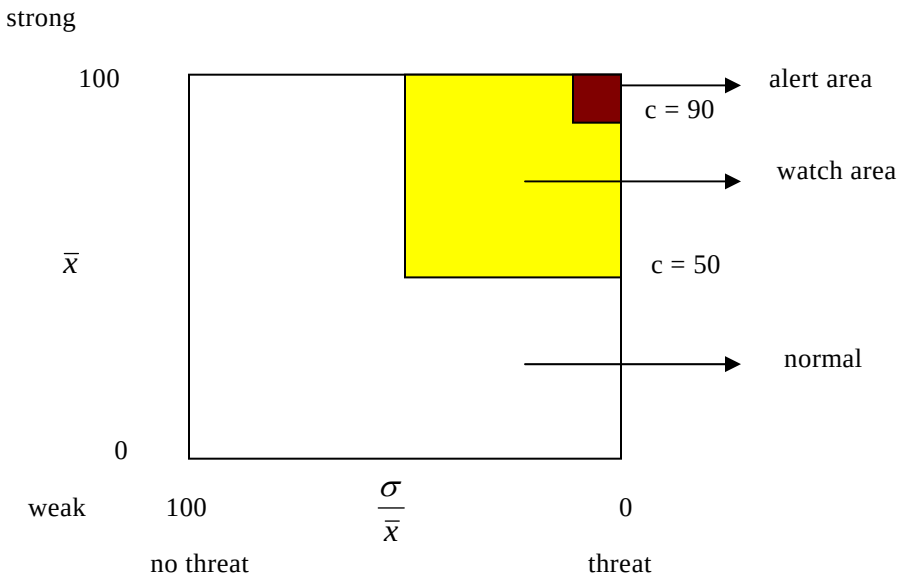


Figure 3.1: Adjusted SWOT Map

In figure 3.1, assume that x is the k-point scale value of view on possibility of terror or some sensitive event observed by experts in the field. $\bar{x}$ is the mean and $\frac{\sigma}{\bar{x}}$ is the coefficient of variation of x's. Note that the c value is the parameter which is the boundary point of each area and should be estimated from the past data including the real event occurred. Then, we can render the black area of the adjusted SWOT map in Figure 3.1 as the alert state and the grey area as the watch state for a risky event.

3.1.6. Procedure of MRSDV model

If the adjusted SWOT map reveals the alert state, and the correlation between x_{1i} and x_{2i}, shows a value higher than 90%, we can render it is a terror state. At that point, MRSDV model must show a steep increase, because the slope of variable x_{1i} is changed from α_i to $\alpha_i\gamma_i$ as follows.

$$y_i = \alpha_0 + \alpha_1 x_{1i} + \beta_1 x_{2i} + \gamma_1 x_{1i} x_{2i} + \varepsilon_i \quad \rightarrow \quad y_i = \alpha_0 + (\alpha_1 + \gamma_1) x_{1i} + \beta_1 + \varepsilon_i$$

And, the organization of observed data for MRSDV model is shown in Table 3.1.

Table 3.1: Organization of observed data for MRSDV model.

$1 \le i : time\ lag \le m$ $1 \le j : observations\ in\ i \le n$	1	2	.....	m
$x_{1i} = \sum_{j=1}^{n} x_{1ij}$	x_{111} : x_{11n}	x_{121} : x_{12n}	.....	x_{1m1} : x_{1mn}
	$\sum_{j=1}^{n} x_{11j}$	$\sum_{j=1}^{n} x_{12j}$		$\sum_{j=1}^{n} x_{1mj}$
$x_{2i} = \sum_{j=1}^{n} x_{2ij}$	x_{211} : x_{21n}	x_{221} : x_{22n}	.....	x_{2m1} : x_{2mn}
	$\sum_{j=1}^{n} x_{21j}$	$\sum_{j=1}^{n} x_{22j}$		$\sum_{j=1}^{n} x_{2mj}$

This procedure for MRSDV model is shown in Figure 3.2 below.

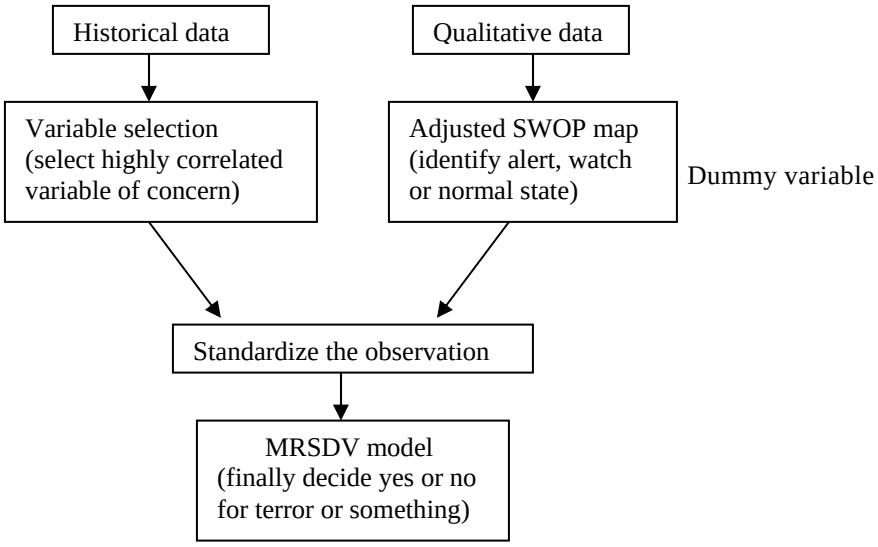


Figure 3.2: Procedure of MRSDV Model

The MRSDV Model depends on the correlation between x_{1i} and x_{2i} such as

$$\text{i) If } corr(x_1, x_2) < c_1, \text{ then } x_2 = 0 \rightarrow y = \alpha_0 + \alpha_1 x_1$$

$$\text{ii) If } corr(x_1, x_2) \geq c_1 \text{ and } \sum_{j=1}^{n} x_{2j} > c_0, \text{ then } x_2 = 1$$

$$\rightarrow y = (\alpha_0 + \beta_1) + (\alpha_1 + \gamma_1) x_1$$

c_0 and c_1 are threshold values determined based on the information from experts and control the effect of qualitative variable (x_2), such that, when $x_2 = 0$, the qualitative variable becomes ineffective; and, when $x_2 = 1$, it becomes effective revealing a spike.

3.2 MRDDV model

As described previously, an ad hoc risk management system (ARMS) is a frame work composed of statistical model and surveyed field data for forecasting. The primary concern in ARMS is dynamically forecasting or predicting a risky state. MRSDV model is insufficient in estimating parameters and lacks fitness test of model and criterion of risk and non-risk states, even though it grasps the spike pattern from actual data. Therefore, we proposed a dependent dummy variable technique, Ordinary Least Square (OLS) method, and fitness test scheme for the proposed model. We also propose a new 'risk plane' for practical use, which is used as threshold criterion of risk.

One of the most important issues in the ARMS is the timing when to watch and warn against risk. The MRSDV model runs between the quantitative independent variable and scaled qualitative independent variable by using dummy variable technique, in which the qualitative variable is surveyed by using k-point scaling method [1, 8, 9]. The proposed MRDDV (Multiple Regression with Dependent Dummy Variable) model, basically, applies the multiple regression method which runs between the quantitative independent variable and the qualitative independent but not dummy variable (it is different from the MRSDV model). It also uses dependent dummy variable technique to set the criterion of risk state and the concept of Ordinary Least Square (OLS) to estimate its parameter [5, 7].

3.2.1. Dependent dummy variable

In this section, we review the concept of dummy variables. There are two methods in dummy variable theory: one is setting independent variable as a dummy variable and another is setting dependent variable as a dummy variable, which gets the value 0 or 1 [2]. In MRDDV model, dependent variable is set as a dummy variable. Assume, for example, simple linear regression model with dependent dummy variable as follows;

$$y_i = \beta_0 + \beta_1 x_{1i} + \varepsilon_i$$

where x stands for personal income (value), and y stands for insurance contract. The value of y is either 0 (for no insurance) or 1 (for insurance).

The expected response is such that;

$$E(y_i) = \beta_0 + \beta_1 x_i = p_i$$

where p_i is the probability for $y_i = 1$.

As shown above, the expected response with respect to dependent dummy variable, represents the probability of occurrence. That means, in this example, that the probability of insurance/ no insurance is changing according to the increase/ decrease

of personal income. The importance of dummy variable is that we can get the outcome such as 0 or 1 and success or failure etc.

3.2.2. Stepwise selection of quantitative variable

To build MRDDV model, firstly, we have to select the variable to fit the response adequately. Hence, we have to find out the variable highly correlated with the response by using stepwise selection technique [5]. For example, the expense-amount of terror group, the number of their phone calls or reservation of air or ground transportation have to be found out as an independent variable highly correlated with the terror occurrence. The reasonable solution for this may be intuitive selection of variables according to experience, or by using variables from similar research already conducted. In statistics theory, there are three kinds of variable selection such as forward selection, backward deletion and stepwise selection. For MRDDV model, this stepwise selection will be adopted to find out the variable that has a high correlation with the outcome of risky event.

3.2.3. Standardization of variable

In case of MRDDV, standardization of all variables is essential because they may have different numerical bases or weights if they are some observed qualities.

Assume, for example, two independent variables that have different bases, in multiple regression, are as follows:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$

where y is sales amount (value), x_1 is advertised amount (value) and x_2 is store size (square feet)

Applying standardization technique to y, x_1, x_2 respectively, we can get the results such as;

$$y_i^* = \frac{y_i - \bar{y}}{s_{y_i}}, \quad x_{1i}^* = \frac{x_{1i} - \bar{x}_1}{s_{1i}}, \quad x_{2i}^* = \frac{x_{2i} - \bar{x}_2}{s_{2i}}$$

where s_y: standard deviation of y, $\bar{y}$: mean of y, s_1: standard deviation of x_1, $\bar{x}_1$: mean of x_1, s_2: standard deviation of x_2, $\bar{x}_2$: mean of x_2

Then, the new regression model can be derived with standardized bases as follows:

$$y_i^* = \beta_0^* + \beta_1^* x_{1i}^* + \beta_2^* x_{2i}^* + \varepsilon_i^*$$

3.2.4 Analytic modelling of MRDDV

The theoretical foundation for MRDDV model with two independent variables will be introduced in this section.

Let's assume the dependent variable as a dummy variable in multiple regression.

Then, it becomes the MRDDV Model, and represented as follows;

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i, \quad y_i = 0,1 \tag{3.2}$$

where,

i : number of data points and $i \geq 2$

y_i : response function represented by dummy variable

x_{1i} : quantitative independent (ith observation in historical data category)

x_{2i} : qualitative independent variable (i*th* mean of experts group)

β_i : parameter to be estimated

e_i : error term

Note that the criterion function for generating y_i in MRDDV, is given as follows.

$$y_i = \begin{cases} 1 & if\ event\ happens \\ 0 & otherwise \end{cases}$$

In MRDDV Model, the dependent variable y can take the value 1 with a probability of success p, or the value 0 with probability of failure $q = 1 - p$, where $0 < p < 1$. This type of variable is called a Bernoulli variable as follows:

$$P(y) = \begin{cases} 1 - p & for \quad y = 0 \\ p & for \quad y = 1 \end{cases} \tag{3.3}$$

Therefore probability density function can be written

$$f(y) = p^{y}(1-p)^{1-y}$$

And the corresponding distribution function is

$$F(y) = \begin{cases} 1 - p & for \quad y = 0 \\ 1 & for \quad y = 1 \end{cases}$$

Expected Value of y, i.e., $E(y) = 0 \times P(y = 0) + 1 \times P(y = 1) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 = p$, is obtained from Formula in Table 3.1 and this formula implies that the expected response of y is the probability that y becomes 1, i.e., $y = 1$. However, estimator $\hat{y} = \hat{E}(y) = b_0 + b_1 x_1 + b_2 x_2$ is not proper to an estimated value $\hat{p}$ of p in some cases, because it violate the definition of probability $0 \le E(y) = p \le 1$. To overcome this violation, Logit Transformation is introduced in section 3.4.

3.2.5. Response function

By setting $E(e_i) = 0$, the formula (3.2) is represented as follows;

$$E(y_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} \tag{3.4}$$

As variable y_i takes only 0 or 1 for its value, the variable y_i is a Bernoulli random variable such as in Table 3.1.

Table 3.2: Probability Distribution of Bernoulli

y_i	probability
0	$P(y_i = 0) = 1 - p_i = q_i$
1	$P(y_i = 1) = p_i$
	$E(y_i) = 0 \times P(y_i = 0) + 1 \times P(y_i = 1) = p_i$

$E(y_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} = p_i$ is obtained from Table1 and this formula implies that, as independent variables x_{1i} and x_{1i} are varying in their domain, the expected response of y_i means the probability that $y_i = 1$.

3.2.6. Parameter estimation of MRDDV model

In MRDDV model, OLS (Ordinary Least Square) estimator for β is unbiased and, furthermore, if the number of observations is large enough, then this model has asymptotically Normal distribution. Parameter estimation for this proposed MRDDV model based on OLS (Theorem A.1 in the Appendix) is as follows;

$$\underline{b} = (X'X)^{-1} X' \underline{y} \tag{3.5}$$

where $\underline{b}$: vector of estimated values of β

X : matrix of observations of independent variable

X' : transpose of matrix X

$(X'X)^{-1}$: inverse matrix of $(X'X)$

$\underline{y}$: matrix of observation on dependent variable

$$X = \begin{vmatrix} 1 & x_{11} & x_{21} \\ 1 & x_{12} & x_{22} \\ : & : & : \\ 1 & x_{1n} & x_{2n} \end{vmatrix} \quad \underline{y} = \begin{vmatrix} 0 \\ 0 \\ : \\ 1 \end{vmatrix} \quad \underline{b} = \begin{vmatrix} b_0 \\ b_1 \\ b_2 \end{vmatrix}$$

note: $\underline{y}$ is a vector of 0s and 1s.

3.2.7. Coefficient of determination on MRDDV model

There are several methods to check the precision of fitness such as MSE (Mean Square Error), F-test and R^2 (Coefficient of determination). In MRDDV model, R^2 is the major criterion to check the goodness of fit of the estimator or predictor where

$$R^2 = \frac{SSR}{SST} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = \frac{\underline{b}' X' \underline{y} - n(\bar{y})^2}{\underline{y}' \underline{y} - n(\bar{y})^2} \tag{3.6}$$

3.3 MRADV Model

MRADV can provide a method to resolve the first two issues with the MRDDV Model, i.e., unavailability of efficient method to achieve i) a better model goodness-of-fit and ii) probabilistic approach to represent the dependent variable.

Firstly, in order to improve the model goodness-of-fit of MRDDV Model, an adjustment made by switching the dependent variable (i.e., y_i), which is represented by a dummy value (e.g., 0 or 1) in MRDDV Model, to the absolute value of the difference between the adjacent current and previous data values, i.e., difference between the *ith* and $(i-1)th$ data points, that is, the criterion function in MRADV as follows.

$$y_i = \left| x_{1i} - x_{1(i-1)} \right|$$

Thus, the criterion function in MRADV that is the absolute value of the forward difference is not only a representation of a scalar quantity of the fluctuation, but it

also reduces the coefficient of determination value $\left(R^2 = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2} \right)$ in MRADV than that in MRDDV model.

Suppose two cases: i) all $y_i = 0$ case and ii) at least one $y_i = 1$ case in MRDDV. Firstly, in case i), $R^2 = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2}$ is undefined. Next, in case ii), MRADV increases R^2 than MRDDV, because R^2 is a correlation (ρ_{xy}) between x and y as follows:

$$R^2 = \frac{SSR}{SST} = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2} = \frac{b_1 \sum (x_i - \bar{x})^2}{\sum (y_i - \bar{y})^2} \qquad \left(\because \ \hat{y} - \bar{y} = b_1 (x - \bar{x}) \right)$$

$$= \left[\frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{y})^2} \right]^2 \frac{\sum (x_i - \bar{x})^2}{\sum (y_i - \bar{y})^2} \qquad \left(\because \ b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \right)$$

$$= \frac{\left[\sum (x_i - \bar{x})(y_i - \bar{y}) \right]^2}{\sum (x_i - \bar{x})(y_i - \bar{y})} = \rho_{xy}$$

Note: the term $\hat{y} - \bar{y} = b_1 (x - \bar{x})$ above is proved in Theorem A.3 in the Appendix.

This result is namely the definition of coefficient of correlation (ρ_{xy}). And, it is obvious that the degree of coefficient of correlation of MRADV is higher than that of MRDDV, because the criterion function of MRADV, $y_i = \left| x_{1i} - x_{1(i-1)} \right|$, is generated from x_i whereas that in MRDDV is given as $y_i = 0 \ or \ 1$.

After the parameter estimation, the regression estimator $\hat{y}_i$ of y_i can be obtained as follows.

$$\hat{y}_i = b_0 + b_1 x_{1i} + b_2 x_{2i} \tag{3.7}$$

where,

i) $\hat{y}_i$: regression estimator of y_i

ii) b_i : estimated parameter for β_i's, respectively.

Next, in order to address and resolve the other issue of the MRDDV Model, i.e., an inefficient way of establishing a criterion for risky and non-risky events, MRADV model employ, the concept of probability, for an event of interest to occur with respect to a criterion. Note that the manipulation of the criterion establishment in the proposed method is as adaptive and manageable as the user needs it to be. The novelty of the proposed method is that it provides an efficient way to facilitate the process to evaluate a probabilistic characteristic based on statistical regression-based estimation of the data, which ultimately will enable the users to make a probabilistic prediction based on the estimated historical trend of the data. Therefore, it will be possible to seamlessly bridge the gap between the data-driven statistical analysis of the data and its probabilistic projection of the trend of the data without making a major assumption, e.g., probability density function, which is a known-major cost in traditional straightforward-probabilistic approaches.

The proposed Probabilistic MRADV is developed in two different approaches, i.e., Observation ($\hat{y}_i$) -Driven Method (ODM) and Input (x_1 and x_2) -Driven Method (IDM). The ODM computes probability of the event of interest based on observation data, i.e., $\hat{y}_i$, with respect to a threshold value chosen from the range of the estimated $\hat{y}_i$, which serves as the solid criterion value queried from the user within the range of $\hat{y}_i$. Likewise, the IDM computes probability based on the input variables, i.e., x_1 and x_2 with respect to the corresponding criterion values that can be computed based on the functional relationship between x_1, x_2 and $\hat{y}_i$, i.e., mapping of the criterion threshold value on the range of $\hat{y}_i$ onto the corresponding values on the range of x_1 and x_2, respectively. Thus, the probabilities measured independently by ODM and IDM against the same set of data are supposed to be in full agreement and this fact will be used for a verification purpose of the correctness of the computation for the probability of the events of interest in this study.

3.3.1 ODM Approach to P_{MRADV}

In order to demonstrate how to derive the probability of an event of interest based on the dependent variable for observation (i.e., ODM) with the proposed adjustment, assume the observation ($\hat{y}_i$) represented by a dependent variable follows the normal distribution with mean of $\mu_{\hat{y}} = \frac{\sum_{i=1}^{n} \hat{y}_i}{n}$ and variance of $\sigma_{\hat{y}}^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \mu_{\hat{y}})^2}{n}$, and thus,

without loss of generality $\hat{y} \sim N\left(\mu_{\hat{y}}, \sigma_{\hat{y}}^2\right)$. Then, the ODM Model for P_{MRADV} can be represented as follows.

$$P_{MRADV} = \int_c^{\infty} f_{odm}(\hat{y}) d\hat{y} = \int_c^{\infty} \frac{1}{\sqrt{2\pi}\sigma_{\hat{y}}} \exp\left\{ -\frac{1}{2}\left(\frac{\hat{y} - \mu_{\hat{y}}}{\sigma_{\hat{y}}} \right)^2 \right\} d\hat{y} \qquad (3.8)$$

where, i) $\mu_{\hat{y}} = \dfrac{\sum_{i=1}^{n} \hat{y}_i}{n}$

ii) $\sigma_{\hat{y}} = \sqrt{\dfrac{\sum_{i=1}^{n} (\hat{y}_i - \mu_{\hat{y}})^2}{n}}$

iii) c is a threshold value

The constant c is a threshold value to be determined by the user in order for P_{MRADV} to manipulate the computation for the probability with respect to the criterion value within the range of $\hat{y}_i$. Figure 3.3 shows a graph of an ODM of P_{MRADV} with respect to $\hat{y}_i$ against a data set, where the x axis represents the range of $\hat{y}_i$ versus the probability density on the y axis.

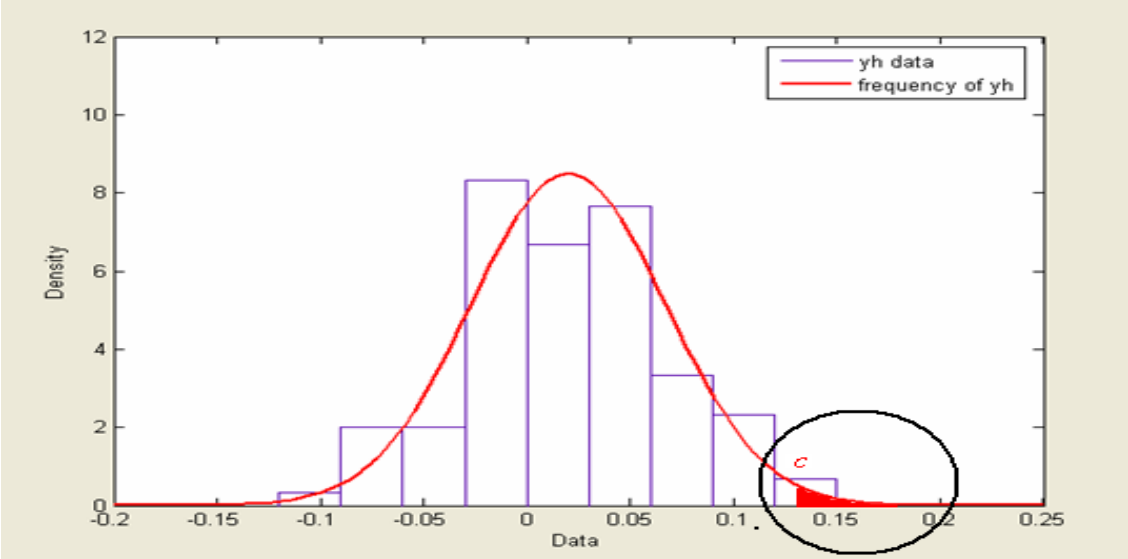


Figure 3.3: ODM of P_{MRADV} with respect to $\hat{y}_i$ as highlighted by the shaded area.

3.3.2 IDM Approach to P_{MRADV}

In the previous section 3.3.1, ODM was proposed to demonstrate how to compute the probability of an event of interest to occur with respect to a threshold value as given by a constant, i.e., *c*, based on the values and the range of a dependent variable for the observation; and in this section it will be shown how to derive the probability based on the values of the independent variables, i.e., x_1 (quantitative independent variable) and x_2 (qualitative independent variable), which is supposed to be in full agreement with the probability derived from $\hat{y}_i$ with respect to the threshold value. It is a reverse-transformation process from the threshold value in the range of $\hat{y}_i$ back on to x_1 and x_2 to find the corresponding values in their range, respectively since $\hat{y}_i$

and x_1, x_2 have been through a common multiple regression and estimated in the same context.

IDM can be distinguished from ODM such that ODM enables the users to compute the probability post to estimation on the multiple regression; while IDM can provide a way to compute the probability prior to estimation on the multiple regression, but in order to share the equivalent criterion obtained from the threshold value, *c* which is supposed to be based on a value in the range of $\hat{y}_i$ and transformed on to the range of x_1, x_2, hence IDM also needs the criterion values to be passed on to it from the estimated dependent variable $\hat{y}_i$.

The equivalence between the probability values obtained through ODM and IDM will be used as a tool to validate the computational correctness by showing they are in full agreement mathematically and statistically.

As ODM-based probability can be expressed by $P(\hat{y} = b_0 + b_1 x_1 + b_2 x_2 \le c)$ under the assumption of normal distribution on $\hat{y}$, the IDM can be expressed by assuming the bivariate normal form of x_1 and x_2, and using the transformation of $\hat{y} = b_0 + b_1 x_1 + b_2 x_2$ as follows.

$$F_{\hat{y}}(c) = P(\hat{y} \le c) = P(b_0 + b_1 x_1 + b_2 x_2 \le c) = \iint_{x_2 \le \frac{c - b_0 - b_1 x_1}{b_2}} f(x_1, x_2) dx_2 dx_1$$

$$\rightarrow \quad P_{MRADV} = 1 - P(b_0 + b_1 x_1 + b_2 x_2 \le c) = 1 - \iint_{x_2 \le \frac{c - b_0 - b_1 x_1}{b_2}} f(x_1, x_2) dx_2 dx_1 \qquad (3.9)$$

where,

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_{x_1}\sigma_{x_2}\sqrt{1-\rho^2}} \exp\left(-\frac{\left(\frac{x_1-\mu_{x_1}}{\sigma_{x_1}}\right)^2 - 2\rho\left(\frac{x_1-\mu_{x_1}}{\sigma_{x_1}}\right)\left(\frac{x_2-\mu_{x_2}}{\sigma_{x_2}}\right) + \left(\frac{x_2-\mu_{x_2}}{\sigma_{x_2}}\right)^2}{2(1-\rho^2)} \right)$$

Note that the threshold value *c* is taken into account in the range of x_2 with respect to x_1 such that $x_2 = \frac{c - b_0 - b_1 x_1}{b_2}$, and vice versa in order to compute P_{MRADV}.

The IDM-based P_{MRADV} can also be expressed by assuming the univariate normal form of x_1 and x_2, and using the fact that $b_0 + b_1 x_1 + b_2 x_2$ is the linear combination of x_1 and x_2. The following property can be obtained from the Theorem A.2 in the Appendix.

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_{x_1} \\ \mu_{x_2} \end{pmatrix}, \begin{pmatrix} \sigma_{x_1}^2 & \rho_{x_1x_2}\sigma_{x_1}\sigma_{x_2} \\ sym & \sigma_{x_2}^2 \end{pmatrix} \right)$$

➔ $$b_0 + b_1 x_1 + b_2 x_2 \sim N\left(b_0 + b_1\mu_{x_1} + b_2\mu_{x_2},\ b_1^2\sigma_{x_1}^2 + b_2^2\sigma_{x_2}^2 + 2b_1b_2\rho_{x_1x_2}\sigma_{x_1}\sigma_{x_2} \right)$$

Let $b_0 + b_1 x_1 + b_2 x_2$ be *U*, then $P(b_0 + b_1 x_1 + b_2 x_2 \le c) = P(U \le c)$.

$$P_{MRADV} = 1 - P(U \le c) = P(U \ge c)$$

$$= \int_c^{\infty} \frac{1}{\sigma_u\sqrt{2\pi}} \exp\left\{ -\frac{1}{2}\left(\frac{u-\mu_u}{\sigma_u}\right)^2 \right\} du \qquad (3.10)$$

where

i) $\mu_u = b_0 + b_1\mu_{x_1} + b_2\mu_{x_2}$

ii) $\sigma_u = \left(b_1^2\sigma_{x_1}^2 + b_2^2\sigma_{x_2}^2 + 2b_1b_2\rho_{x_1x_2}\sigma_{x_1}\sigma_{x_2}\right)^{1/2}$

iii) c is a threshold value

The bivariate normal form in copula based method [16] and the univariate normal form in ODM-based method are equivalent in terms of their probabilities by the probabilistic property $P(b_0 + b_1x_1 + b_2x_2 \le c)$. The bivariate normal form in formula (3.8), provides an efficient way to visually plot the graph of joint probability of x_1 and x_2.

3.3.3 Equivalence between ODM and IDM

From Equations (3.9) and (3.10), it is shown that ODM and IDM approaches compute equivalent results for P_{MRADV}.

From Equation **(**3.9), the following was obtained

$$P_{MRADV} = \int_c^{\infty} f_{odm}(\hat{y})d\hat{y} = \int_c^{\infty} \frac{1}{\sqrt{2\pi}\sigma_{\hat{y}}} \exp\left\{-\frac{1}{2}\left(\frac{\hat{y}-\mu_{\hat{y}}}{\sigma_{\hat{y}}}\right)^2\right\}d\hat{y}$$

and from Equation (3.10), an equivalent was obtained as follows.

$$P_{MRADV} = P\left(\hat{y} = b_0 + b_1x_1 + b_2x_2 \ge c\right) = P(U \ge c) = \int_c^{\infty} \frac{1}{\sigma_u\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{u-\mu_u}{\sigma_u}\right)^2\right\}du$$

Notice that both $\hat{y}$ in equation (3.9) and u in equation (3.10) represent $b_0 + b_1 x_1 + b_2 x_2$ as was shown in the previous sections.

In general, there is no known effective way to compute an integral in the form of

$$\int_{-\infty}^{k} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}h^2\right\} dh$$

that is needed to compute P_{MRADV} , because the anti-derivative of

$$\exp\left\{-\frac{1}{2}h^2\right\}$$

can not be readily computed by employing any standard algebraic method. Hence, in this study, the table of the approximate values of this integral for various values are used as given in Table A.1 in the Appendix, and if needed more precise values can be approximated by using the method of interpolation [5].

3.3.4 Relation with Copula

In order to compute joint probability of x_1 and x_2, we employ Copulas.

A copula is a function $C : [0,1]^2 \to [0,1]$ which satisfies the following conditions [10]:

i) for every u , v in $[0,1]$, $C(u,0) = 0 = C(0,v)$, and $C(u,1) = u$ and $C(1,v) = v$

ii) for every u_1, u_2, v_1, v_2 in $[0,1]$ such that $u_1 \le u_2$ and $v_1 \le v_2$,

$C(u_2,v_2) - C(u_2,v_1) - C(u_1,v_2) + C(u_1,v_1) \ge 0$.

Based on the Sklar's Theorem [10], the copula model of interest can be derived as follows.

Let X and Y be random variables with joint distribution function H and marginal distribution function F and G, respectively. Then, there exists a copula C such that

$$H(x,y) = C(F(x), G(y)), \text{ for all } x, y \text{ in } \Re \tag{3.11}$$

Conversely, given a copula C and distributions F and G, the function H defined by equation (3.11) is a bivariate distribution with margins F and G.

In particular, if X and Y are extended real valued random variables, defined on a common probability space, with individual distribution F_X and F_Y and joint distribution $F_{X,Y}$, then there is a copula $C_{X,Y}$ such that $F_{X,Y}(u,v) = C_{X,Y}(F_X(u), F_Y(v))$. If F_X and F_Y are continuous, $C_{X,Y}$ is unique. It is referred to $C_{X,Y}$ as a copula of X and Y. The copula of two random variables thus reveals their dependence structure.

In addition, Gaussian Copula is defined as the following copula:

$$C(u,v) = G\left(\Phi^{-1}(u), \Phi^{-1}(v) \mid \rho\right)$$

$$= \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\Phi^{-1}(u)\right)^2 + \left(\Phi^{-1}(v)\right)^2 - 2\rho\left(\Phi^{-1}(u)\right)\left(\Phi^{-1}(v)\right)\right]\right\} \tag{3.12}$$

where

$$u, v \sim N(0,1),\ \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{u^2}{2}\right) du : cdf \quad of \quad N(0,1) \quad -1 < \rho < 1.$$

The P_{MRADV} model in consideration to compare with Gaussian Copula is as follows.

$$P_{MRADV} = \int_{c}^{\infty} \frac{1}{\sqrt{2\pi}\sigma_{\hat{y}}} \exp\left\{-\frac{1}{2}\left(\frac{\hat{y} - \mu_{\hat{y}}}{\sigma_{\hat{y}}}\right)^2\right\} d\hat{y} = 1 - N(c)$$

$$\text{Copula} = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\Phi^{-1}(u)\right)^2 + \left(\Phi^{-1}(v)\right)^2 - 2\rho\left(\Phi^{-1}(u)\right)\left(\Phi^{-1}(v)\right)\right]\right\}$$

$$= \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[N^{-2}(u) + N^{-2}(v) - 2\rho N^{-1}(u) N^{-1}(v)\right]\right\}$$

Note that in the above P_{MRADV}, $N(c)$ can be calculated by using Table A.1 in the Appendices, which is the calculation of standard normal distribution.

In order to make the copula be comparable with P_{MRADV} in the context of IDM-based method which is equivalent to ODM-based, the threshold value c is transformed into the context of copula such that the variables u and v in the copula are transformed as follows.

$$b_0 + b_1 u + b_2 v = c$$

3.4 Logit Transformation-based MRDDV Model

In this section, firstly, MRDDV Model is reviewed and Logit Transformation is introduced. Next, original MRDDV Model is transformed in Logit form, namely Logit Transformation-based MRDDV Model, and compared with original MRDDV Model in term of risk probability.

Let's assume the dependent variable as a dummy variable in multiple regression. Then, it becomes the MRDDV Model in equation (3.2).

Estimator $\hat{y} = \hat{E}(y) = b_0 + b_1 x_1 + b_2 x_2$ is not proper to an estimated value $\hat{p}$ of p in some cases, because it violate the definition of probability $0 \le E(y) = p \le 1$. To overcome this violation, Logit Transformation is introduced in next section.

3.4.1 Logit Transformation

To overcome the violation of probability definition to estimate $E(y) = P(y = 1)$ in original MRDDV Model, we employed Logit Transformation defined as follows:

Definition: Logit Transformation

$$\text{Logit}(p) = \begin{cases} \log \dfrac{p}{1-p} & , \ if \ 0 \le p \le 1 \\ undefined & , \ otherwise \end{cases} \qquad (3.13)$$

Applying this transformation to original MRDDV Model, we specify a variable *p* above, as a probability of dependant dummy variable (y_i) takes 1 in original MRDDV Model.

3.4.2 Logit Transformation-based MRDDV Model

Original MRDDV Model can be expressed in terms of Logit as follows:

$$\log_e\left[\frac{P(y_i=1)}{1-P(y_i=1)}\right]=\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+e_i \tag{3.14}$$

where $-\infty<\log_e\left[\frac{P(y_i=1)}{1-P(y_i=1)}\right]<\infty$

And, it can be arranged in terms of $P(y_i=1)=p_i$ as follows:

$$P(y_i=1)=\frac{\exp(\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+e_i)}{1+\exp(\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+e_i)}$$

Then, it becomes the Logit Transformation-based MRDDV Model as follows:

$$E(y_i)=\frac{\exp(\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+e_i)}{1+\exp(\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+e_i)} \tag{3.15}$$

where,

i : number of data points and $i \geq 2$

y_i : response function represented by dummy variable i.e. 0 or 1

x_{1i} : quantitative independent (ith observation in historical data category)

x_{2i} : qualitative independent variable (i*th* mean of experts group)

β_i : parameter to be estimated

e_i : residuals

3.4.3 Parameter Estimation

In Logit Transformation-based MRDDV Model, the parameter is estimated by MLE (Maximum Likelihood Estimation) as follows:

$f(y) = p^y (1-p)^{1-y}$ (probability density function)

➔ $L(\beta_0, \beta_1, \beta_2) = \prod_{i=1}^{n} f(y_i) = \prod_{i=1}^{n} p^{y_i} (1-p)^{1-y_i}$ (likelihood function)

➔ $\log L = \log_e \prod_{i=1}^{n} f(y_i) = \log_e \prod_{i=1}^{n} p^{y_i} (1-p)^{1-y_i} = \sum_{i=1}^{n} y_i \log_e p + \sum_{i=1}^{n} (1-y_i) \log_e (1-p)$

$= \sum_{i=1}^{n} \left[y_i \log_e \left(\frac{p}{1-p} \right) \right] + \sum_{i=1}^{n} \log_e (1-p)$ (logarithm)

note) $\log_e \left(\frac{p}{1-p} \right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ and $(1-p) = \left[1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2) \right]^{-1}$

$$\rightarrow \log L = \sum_{i=1}^{n} y_i \left(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}\right) + \sum_{i=1}^{n} \log_e \left[1 + \exp\left(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}\right)\right]$$

Then, we can get the estimator $b_i's$ by solving the equations by equating the partial differentials to zero:

$$\frac{\partial \log_e L}{\partial \beta_0} = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \frac{\exp\left(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}\right)}{1 + \exp\left(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}\right)} = 0$$

$$\frac{\partial \log_e L}{\partial \beta_1} = \sum_{i=1}^{n} y_i x_{1i} - \sum_{i=1}^{n} \frac{x_{1i} \exp\left(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}\right)}{1 + \exp\left(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}\right)} = 0$$

$$\frac{\partial \log_e L}{\partial \beta_2} = \sum_{i=1}^{n} y_i x_{2i} - \sum_{i=1}^{n} \frac{x_{2i} \exp\left(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}\right)}{1 + \exp\left(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}\right)} = 0$$

$b_i's$ can be computed using by SAS (Statistical Analysis System) package [32], we employ those results from SAS in this study.

After parameter estimation in Logit Transformation-based MRDDV Model, the estimator $\hat{E}(y_i) = \hat{P}(y_i = 1)$ is provided as follows:

$$\hat{E}\left(y_i\right) = \frac{\exp\left(b_0 + b_1 x_{1i} + b_2 x_{2i}\right)}{1 + \exp\left(b_0 + b_1 x_{1i} + b_2 x_{2i}\right)} \tag{3.16}$$

where,

i) $\hat{E}(y_i)$ is a regression estimator of $E(y_i)$

ii) b_i 's are the results of parameter estimation for β_i 's, respectively

Formula (3.15) in previous section, is a risk probability of Logit Transformation-based MRDDV Model, $\hat{E}(y_i) = \frac{\exp(b_0 + b_1 x_{1i} + b_2 x_{2i})}{1 + \exp(b_0 + b_1 x_{1i} + b_2 x_{2i})} = \hat{p}_i$, and it runs between 0 and 1.

CHAPTER IV

SIMULATIONS

4.1 Evaluation of MRSDV Model

The theoretical model presented in the previous sections provides a basic understanding of the performance impact of ad hoc risk management system. However, the model uses somewhat unidentified c value and observation with outlier. In this section, we use an artificial data set to evaluate MRSDV model.

We simulate the MRSDV model with the simplest form;

$$y = \alpha_0 + \alpha_1 x_1 + \beta_1 x_2 + \gamma_1 x_1 x_2 + e,$$

where x_1 is quantitative variable and x_2 is qualitative variable

For our simulation, let us assume that '*E*' be a critical event to be predicted using the above model. Let us further assume that qualitative variable represents a property *f(t)* and the qualitative variable represent the property *g(t)*. Then *f(t)* is an activity

associated to the event *E* and *g(t)* is qualitative knowledge associated to the event *E*. We also assume the data shown in Tables 4.1 and 4.2 for our analysis.

Firstly, let's think about the data shown in Table 4.1 which are randomly generated and assume x_1 = *f(t)* shown in Table 4.1 and which represent the stable state till time 7, but in time 8, it increases 3 times versus previous time.

Table 4.1: Sample Data Set of Quantitative Variable

time	1	2	3	4	5	6	7	8	9	10
f(t)	156	207	200	176	111	193	190	666	250	184

The fitted equation is represented such as;

$$y = 55 + 41x_1 \text{ ,} \qquad \text{before standardization}$$

$$= -1.03 + 0.230x_1 \text{ ,} \qquad \text{after standardization} \qquad (4.1)$$

Next, let's also think about the data shown in Table 4.2 and assume x_2 be the knowledge represented by means of k-point scaling method with k=100 (that is equivalent with percentage) from 10 experts.

Table 4.2: Sample Data Set of Qualitative Variable

time	1	2	3	4	5	6	7	8	9	10
mean of k-point scaling ($\bar{x}$)	6.9	3.0	7.4	6.4	5.4	4.4	6.6	90.6	10.3	5.7
variance of k-point scaling (s_x)	8.05	2.38	9.68	6.88	3.36	5.16	8.30	3.23	10.21	6.43

This data set in Table 4.2 plots on the adjusted SWOT map, is shown in Figure 4.1. In this SWOT map, we can see 1 to 7 is in stable state but time 8 is in alert state. It is a symptom of the occurrence of event "*E*".

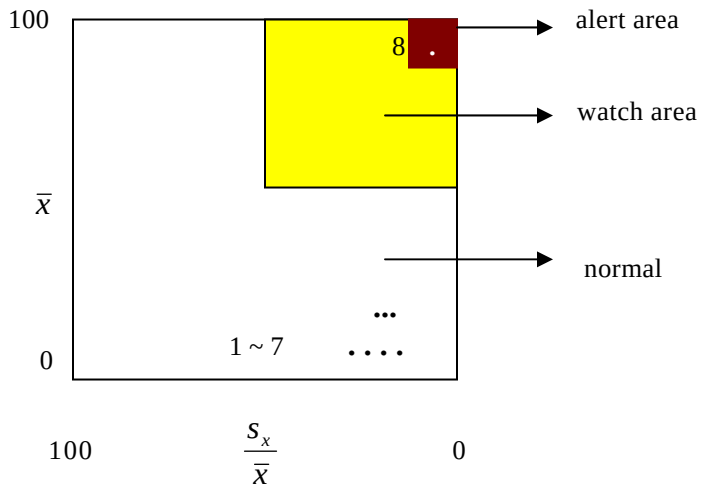


Figure 4.1 Interpretation of SWOT Map

Now, we standardize the observation of quantitative variable and calculate the correlation between two variables to apply qualitative variable as a dummy form. The result of this calculation is shown below;

$$y = (\alpha_0 + \beta_1) + \alpha_1 x_1 = 0.794 + 0.230 x_1, \qquad \text{if corr } (x_1, x_2) < 0.9 \tag{4.2}$$

$$= (\alpha_0 + \beta_1) + (\alpha_1 + \gamma_1) x_1 = 0.794 + 0.728 x_1, \qquad \text{if corr } (x_1, x_2) \geq 0.9 \tag{4.3}$$

Finally, the comparison of equations (4.1), (4.2) and (4.3) is shown in Figure 4.2.

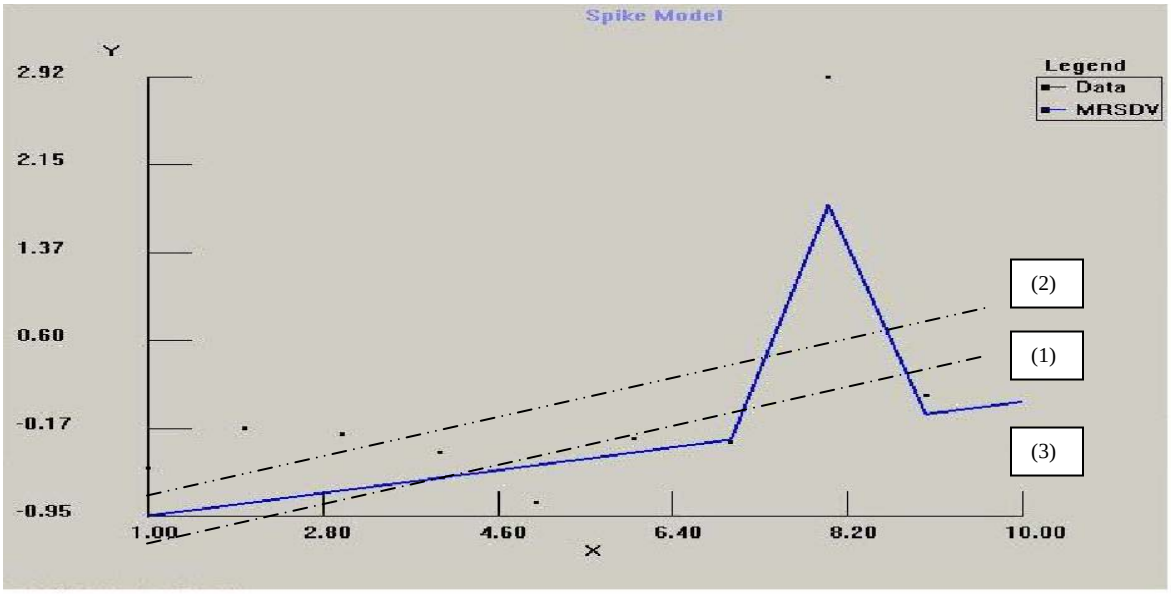


Figure 4.2: Comparison of Multi-Linear Model and MRSDV Model

(2): REGRESSION with historical data and dummy data lowly correlated

(1): REGRESSION with historical data only

(3): MRSDV model with historical data and dummy data highly correlated

As seen in Figure 4.2 multiple regression (2) is not capable of predicting the critical point t = 8, where as when the qualitative term is introduced, the jump at t = 8 becomes obvious. The MRSDV model is very sensitive because the knowledge which is represented as a qualitative variable surveyed by means of k-point scaling method, is reflected as a dummy variable to MRSDV.

The MRSDV Model depends on the correlation between x_{1i} and x_{2i} such as

i) If $corr(x_1, x_2) < c_1$, then $x_2 = 0$ ➔ $y = \alpha_0 + \alpha_1 x_1 + e$

ii) If $corr(x_1, x_2) \geq c_1$ and $\sum_{j=1}^{n} x_{2j} > c_0$, then $x_2 = 1$

➔ $y = (\alpha_0 + \beta_1) + (\alpha_1 + \gamma_1) x_1 + e$

c_0 and c_1 are threshold values determined based on the information from experts and control the effect of qualitative variable (x_2), such that, when $x_2 = 0$, the qualitative variable becomes ineffective; and, when $x_2 = 1$, it becomes effective revealing a spike.

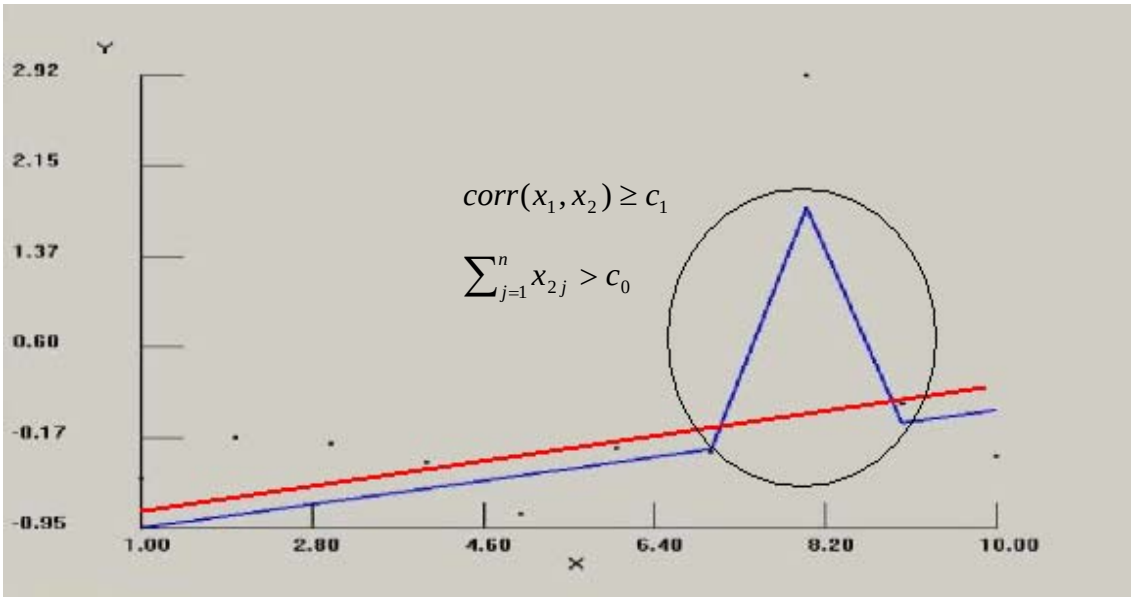


Figure 4.3: Example plotting of results; MRSDV: blue line, SR: red line

The key advantage of MRSDV model is practical to estimate and/ or predict against the risk using not only historical data but also the knowledge and intelligence of experts in field. We also provide the theoretical model to merge these knowledge and intelligence with historical data by means of dummy variable. The modelling effort also shows, if we accumulate the related data, the parameter of MRSDV model can become stable. This indicates the proposed MRSDV model is a powerful tool to predict unforeseen events. The simulation study also shows the superiority of MRSDV model than traditional multiple regression only with historical data. The drawbacks of MRSDV model are:

i) MRSDV Model lacks criterion of risk/ non-risk: MRSDV Model can catch the spike pattern under the control of qualitative variable without any definite risk criterion.

ii) MRSDV Model has no available dependent variable (y_i) to use in its model: actually, in risky event or severe disaster, there is no actual data available as a response.

iii) MRSDV Model is inefficient in its parameter estimation: as shown in MRSDV Model, parameters are mutilated from the original after estimation such as from α_1 to $\alpha_1\gamma_1$, which may cause a deviation of the estimator from the true result. This may eventually result in an error on estimator of response function ($\hat{y}_i$), and further result in an inaccurate model fitness.

4.2. Evaluation of MRDDV Model

The theoretical model presented in Chapter III provides a basic understanding of the performance impact of ad-hoc risk management system by using MRDDV model. In this section, a simple form of MRDDV model is used to evaluate it by simulation. The following model is used:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i, \quad y_i = 0,1 \tag{4.4}$$

where i : number of observations

y_i : risk function with response 0 or 1

x_{1i} : quantitative variable

x_{2i} : qualitative variable

β_i : coefficient of each variable

e_i : error

There are two kinds of simulation such as i) conceptual approach between response function between independent variables, and ii) practical approach with actual industry data with respect to response function.

Firstly, we have two assumptions for conceptual approach such as: 1) x_{1i} is the independent quantitative variable selected through the stepwise procedure. 2) 100 is selected for k value of k-scaling in independent qualitative variable. There are two distinct cases that occur namely “hit case” and “no hit case”. In the first case, the regression line meets the risk plane and in the second case, it does not. The data set shown in Table 4.3 and Table 4.4 are randomly generated for this simulation. We also discuss how the results can be used for prediction.

4.2.1. Conceptual Approach

4.2.1.1 Hit case

A. General estimation of hit case

In Table 4.3, we represent the observation for response function (y_i **,** dependent dummy variable) , running expenses (x_{1i} **,** independent quantitative variable), field

data (x_{2i} , independent qualitative variable) and the standardization results (x_{1i}^{*}, x_{2i}^{*}) of two independent variables, respectively.

Table 4.3: Sample Data Set for Hit Case

i	Running Expenses (x_{1i})	Stand-ardized (x_{1i}^{*})	Field Data (x_{2i})	Stand-ardized (x_{2i}^{*})	Response Function (y_i)	Probability ($\hat{p}_i$)
1	156	-.52	11.4	-.65	0	.28
2	207	-.14	30.2	-.16	0	.38
3	200	-.19	22.2	-.18	0	.37
4	176	-.37	18.4	-.35	0	.33
5	111	-.86	8.0	-.79	0	.22
6	193	-.24	19.0	-.32	0	.35
7	190	-.27	16.4	-.43	0	.33
8	567	2.59	85.6	2.55	1	1.00

The estimation for β_i by the method of OLS (Ordinary Least Square) and R^2 (coefficient of determination) are

$$b = \begin{vmatrix} b_0 \\ b_1 \\ b_2 \end{vmatrix} = (X'X)^{-1}X'y = \begin{vmatrix} .41 \\ .12 \\ .11 \end{vmatrix}$$

$$R^2 = \frac{SSR}{SST} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = \frac{\underline{b}' X' \underline{y} - n(\bar{y})^2}{\underline{y}' \underline{y} - n(\bar{y})^2} = 0.67$$

The coefficient of determination R^2 ranges from 0 to 1. $R^2 = 1$ means the model is a perfect fit. In this simulation, MRDDV model fits data (historical and field data) at the level 67% since $R^2 = .67$. Also, the estimated response function is

$$E(y) = .41 + .12x_1^* + .11x_2^* \qquad (4.5)$$

The result is shown in Figure 4.4. In this case, the expected response hits the risk plane.

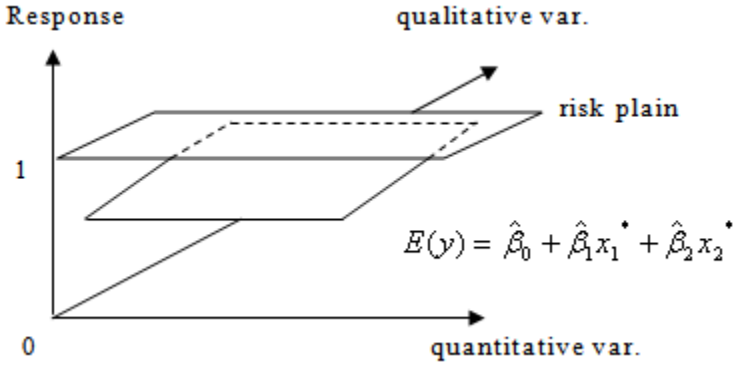


Figure 4.4: Hit Case on MRDDV Model

B. Prediction based on hit case

To use, the model for prediction, we need conditions that can be used to predict anticipated risk. The condition that we propose is "the response function crossing the risk plane". Therefore, the pair (x_{1i}, x_{2i}) for which the response function is greater than or equal to 1, indicates a risky event. For example, if the input for Formula (4.5) is 2.84 for x_1^* and 2.74 for x_2^* which is the value after standardization of 600 for x_{1i} and 90 for x_{2i}, respectively, then the response is 1.05 and exceeds 1. The result is shown in Figure 4.5. This case can be predicted as a risky event.

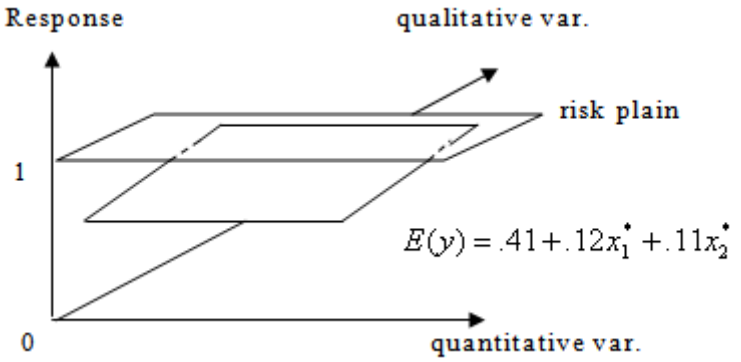


Figure 4.5: Prediction based on Hit Case

4.2.1.2 No hit case

A. General estimation of no hit case

In Table 4.4, we represent the observations of a “no hit” case. It would be a special case of hit case with the same data from time1 to time7 of hit case. In this case, $\underline{y}$ is 0 vector, $\underline{b}$ by the method of OLS (Ordinary Least Square) should be 0 vector. So, the response function is

$$E(y) = 0 \tag{4.6}$$

Table 4.4: Sample Data Set for No Hit Case

i	Running Expenses (x_{1i})	Field Data (x_{2i})	Response Function (y_i)
1	156	11.4	0
2	207	30.2	0
3	200	22.2	0
4	176	18.4	0
5	111	8.0	0
6	193	19.0	0
7	190	16.4	0

This result is shown in Figure 4.6. The response function in this case cannot meet the risk plane.

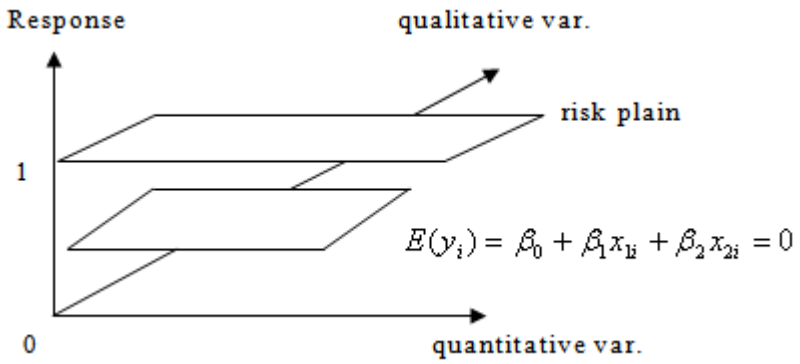


Figure 4.6: No Hit Case on MRDDV Model

B. Seeding for no hit case

In the "no hit case", the conditions of risk cannot be applied as we described earlier because that case requires one observation with $y_i = 1$. To overcome this problem, we can inject the artificial data set including the value 1 for response function (y_i) as a seed to activate the model. For example, if we inject the data set of time 8 of hit case as the 8^{th} data set of no hit case, then we have exactly same simulation result with the hit case. Furthermore, this technique enables the user to define the risky event injecting both independent variables (historical data and field data) artificially if there is no actual risky event data.

4.2.2 Practical Approach

In this section, MRDDV Model is evaluated via simulation with actual industry data on the term of estimated response function ($\hat{y}$).

The data sets are from the financial sector as Dow Jones Industrial Average as quantitative variable (x_1) and Dow Jones-AIG Commodity Index as a qualitative independent variable (x_2). The data sets are based on the daily closing price, and they are over the time period from Feb. 1, 2007 to Apr. 30, 2007.

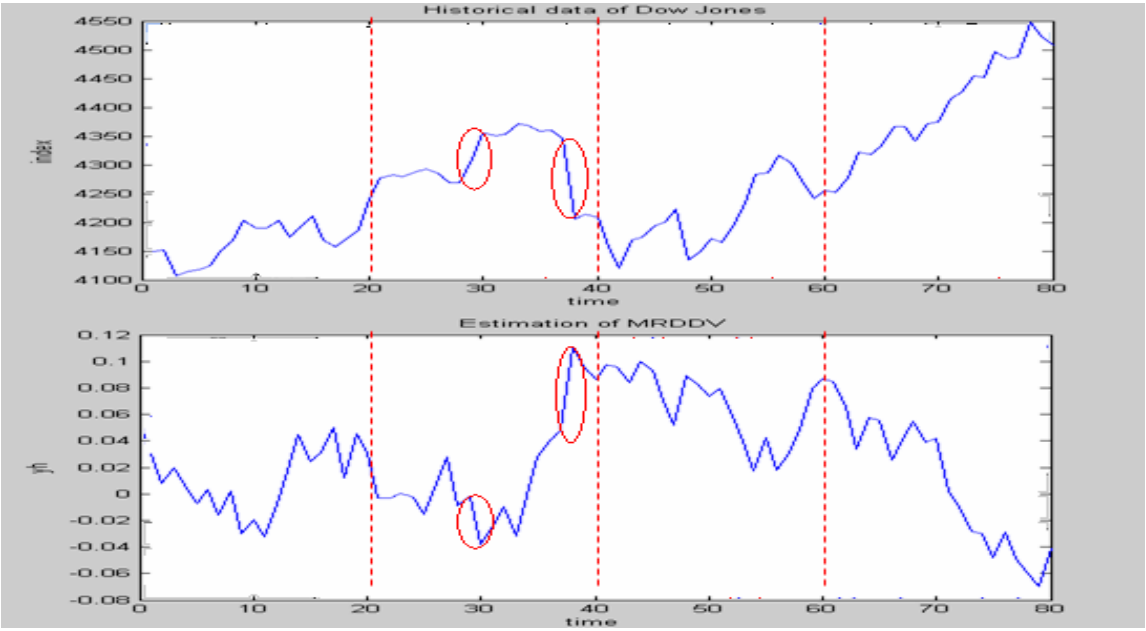


Figure 4.7: Comparison between historical data (Dow Jones) with risk in MRDDV

Figure 4.7 shows the comparison between historical data in the top figure and the estimated response of MRDDV model in the bottom figure. The estimated response (y_h) represents the risks by tracing effectively the historical data in reverse way as shown in the highlighted by oval. But it lacks of probabilistic measure because it shows negative value in some intervals.

4.3 Experimental Study on MRADV Model

In this section, the efficiency and effectiveness of the proposed ODM and IDM-based P_{MRADV} computations will be demonstrated against an extensive set of data in the financial sector through extensive experimental simulations. Note that since ODM and IDM-based approaches are equivalent, ODM-based is chosen for the comparison for more efficient computation purpose. The primary purpose of using the data from the financial sector is that the trend of the data in the field is known to be one of the data sets that has historically exercised good examples of the turbulent market, which is the kind of risky events of interest in this study.

P_{MRADV} is evaluated with respect to various threshold values in its full possible given range under the assumption of Gaussian Distribution. The primary simulation tool to be used is Matlab. For the ODM-based case as shown in Figure (3.3), P_{MRADV} is plotted versus the estimated observation ($\hat{y}$) values; and for the IDM-based case, P_{MRADV} is plotted versus the inputs x_1 and x_2 values as shown in Figure (4.8).

The data sets from the financial sector to be used for the extensive experimental simulations are collected from [29, 30, 31] as follows:

- Dow Jones Industrial Average as quantitative independent variable (x_1)
- Dow Jones-AIG Commodity Index as a qualitative independent variable (x_2)

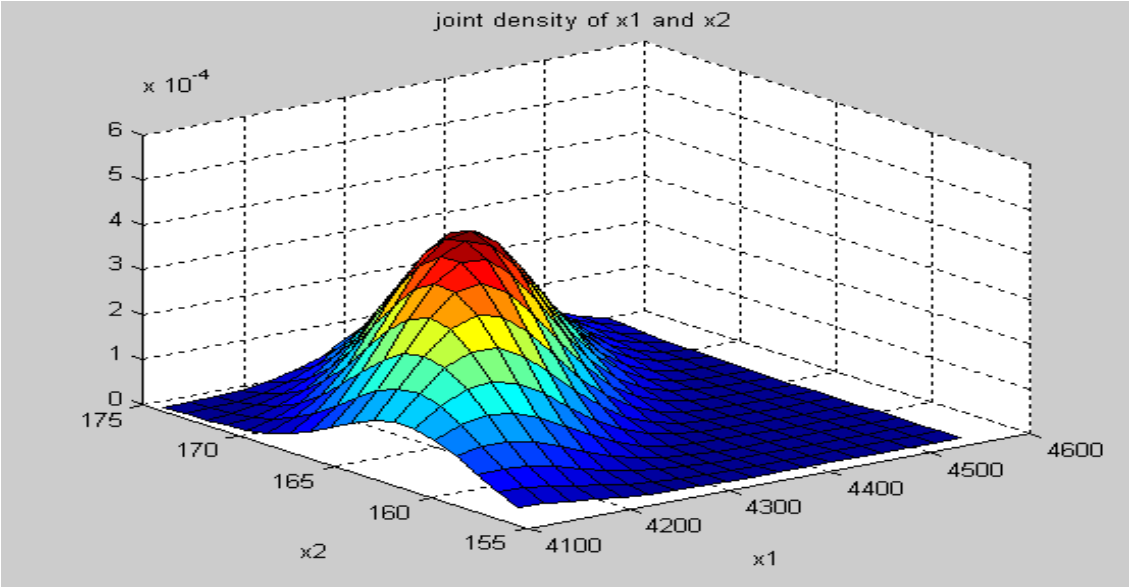


Figure 4.8: A Joint Density of x_1(DJI) and x_2(DJC) by Bivariate Normal Distribution

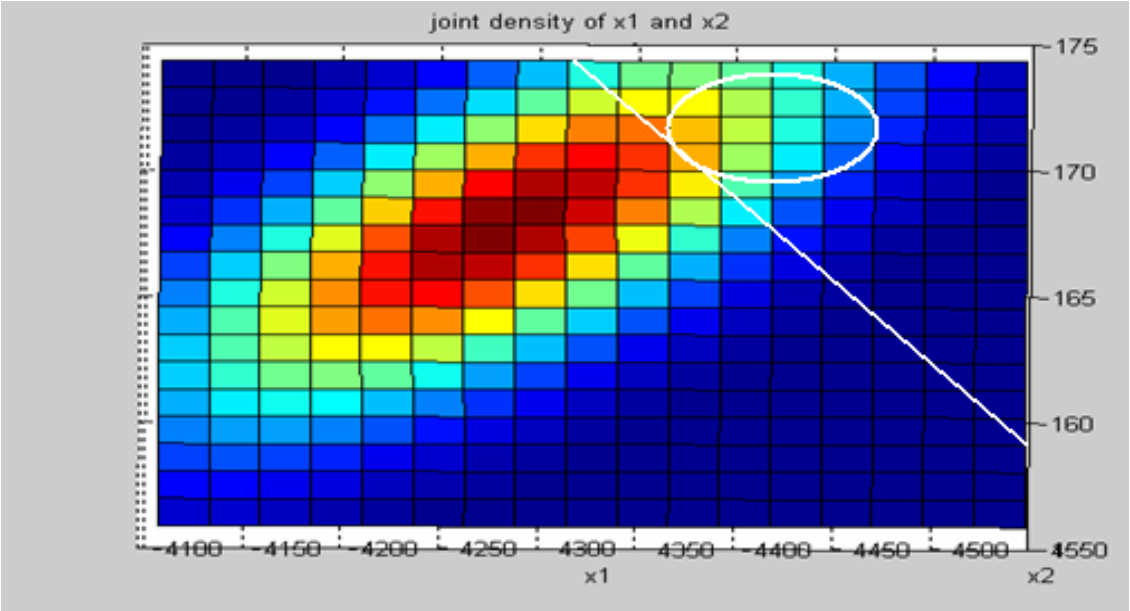


Figure 4.9: Contour Diagram of x_1 and x_2.

Figure 4.8 shows a joint density of x_1 and x_2 depicted in probabilistic way by IDM-based P_{MRADV} model that is given below:

$$P_{MRADV} = 1 - P(b_0 + b_1 x_1 + b_2 x_2 \le c) \quad = 1 - \iint_{x_2 \le \frac{c - b_0 - b_1 x_1}{b_2}} f(x_1, x_2) dx_2 dx_1$$

In Figure 4.9, the contour diagram of Figure 4.8 is shown and the highlighted straight line represented by $x_2 = \dfrac{c - b_0 - b_1 x_1}{b_2}$ is an user's input by deciding the level of threshold value *c* to determine the range for x_2; and the area highlighted by the oval is a spatial representation of P_{MRADV} with respect to x_2. Once threshold value *c* is decided, the risk probability can be computed efficiently by ODM-based P_{MRADV} as follows:

$$P_{MRADV} = 1 - P(U \le c) = P(U \ge c) \quad = \int_c^{\infty} \frac{1}{\sigma_u \sqrt{2\pi}} \exp\left\{ -\frac{1}{2} \left(\frac{u - \mu_u}{\sigma_u} \right)^2 \right\} du$$

Due to the fact both IDM and ODM-based methods reach equivalent probabilistic values. The probability can be simply provided by using threshold value *c* as an input constant in the Table A.1 in the Appendix. To compute joint probability of x_1 and x_2, we either employ conventional Copula method or other joint probability approaches.

4.4 Simulation with Logit Transformation-based MRDDV Model

In this section, comparative study between original MRDDV Model and Logit Transformation-based MRDDV Model is conducted on the term of risk probability. The probabilistic model presented in the previous section, provide a basic

understanding of the risk probability and the models investigated in this study are listed as follows:

Logit Transformation-based MRDDV Model

$$\hat{E}(y_i) = \frac{\exp(b_0 + b_1 x_{1i} + b_2 x_{2i})}{1 + \exp(b_0 + b_1 x_{1i} + b_2 x_{2i})} \tag{4.9}$$

where

$$E(y) = p(y = 1) \quad \text{where} \quad y = \begin{cases} 1 & if \ |x_{1i} - x_{1(i-1)}| \geq 300 \\ 0 & if \ |x_{1i} - x_{1(i-1)}| < 300 \end{cases}$$

The data sets from the financial sector in Table A.2, Appendix are collected from [13] with the properties as follows:

- Dow Jones Industrial Average as quantitative independent variable (x_1)
- Dow Jones-AIG Commodity Index and 10-years Treasury Note as a qualitative independent variable (x_2), for two different cases, respectively.

Table 4.5 Summary of Financial Data Set

Index \ Statistics	Mean	Variance
Dow Jones Industrial Average	4.2723e+003	1.2008e+004
Dow Jones-AIG Commodity Index	167.2807	25.4332
10-years Treasury Note	4.6784	0.0106

The data sets are based on the daily closing price, respectively, and they are over the time period from Jan. 1, 2007 to Dec. 31, 2007. The main statistics of those data sets are summarized in Table 4.5. In this simulation, Matlab is used as a major tool to evaluate Logit Transformation-based MRDDV Model and SAS statistical package is either used to estimate the parameter and results is shown in Table 4.6.

Table 4.6: Financial Parameter Estimation

finance parameter estimation 2
22:06 Thursday, February 7, 2008

The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-21.9176	12.0023	3.3347	0.0678
dji	1	-0.00018	0.000785	0.0539	0.8164
djc	1	0.1207	0.0616	3.8388	0.0501

22:06 Thursday, February 7, 2008

The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.0945	10.8600	0.0102	0.9197
dji	1	0.000524	0.000741	0.4991	0.4799
trx	1	-2.0103	1.0623	3.5814	0.0584

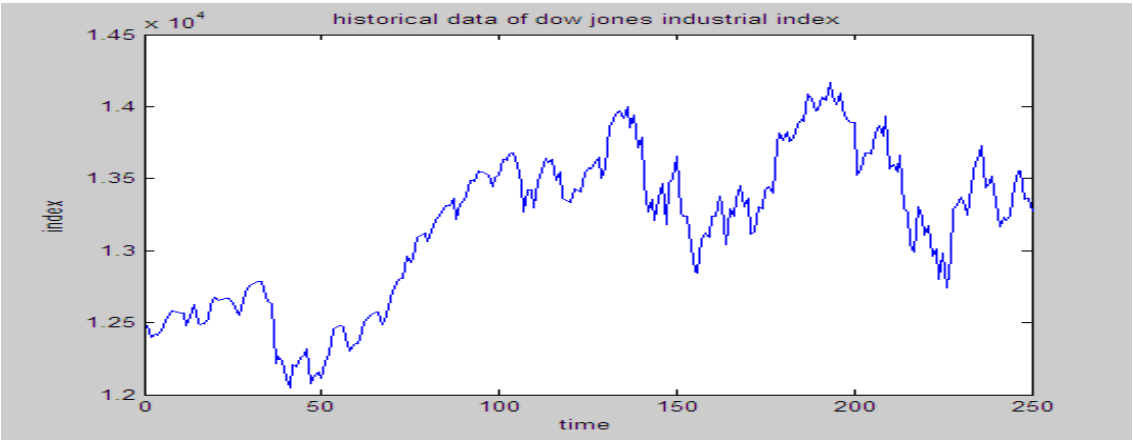


Figure 4.10: Historical data of Dow Jones Industrial Index (DJI)

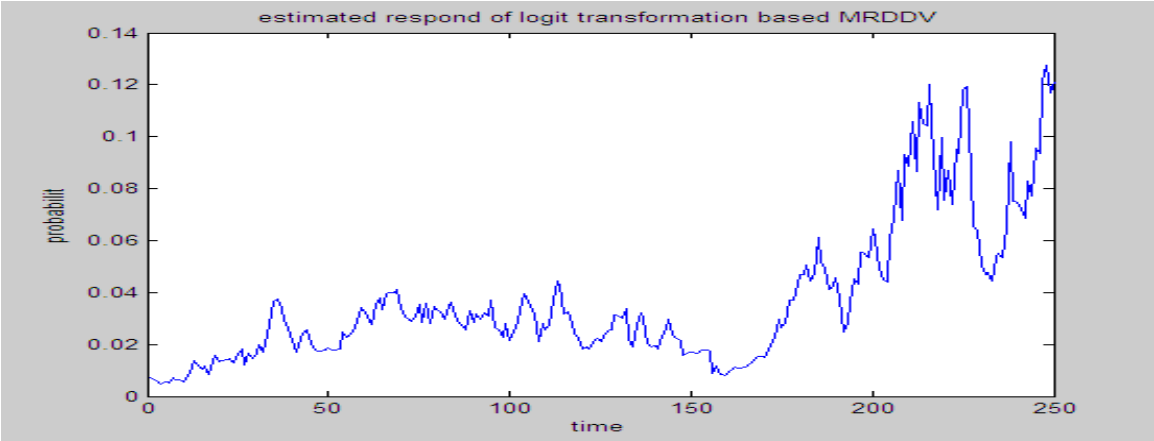


Figure 4.11: Estimated response of Logit Transformation-based MRDDV – DJI*DJC

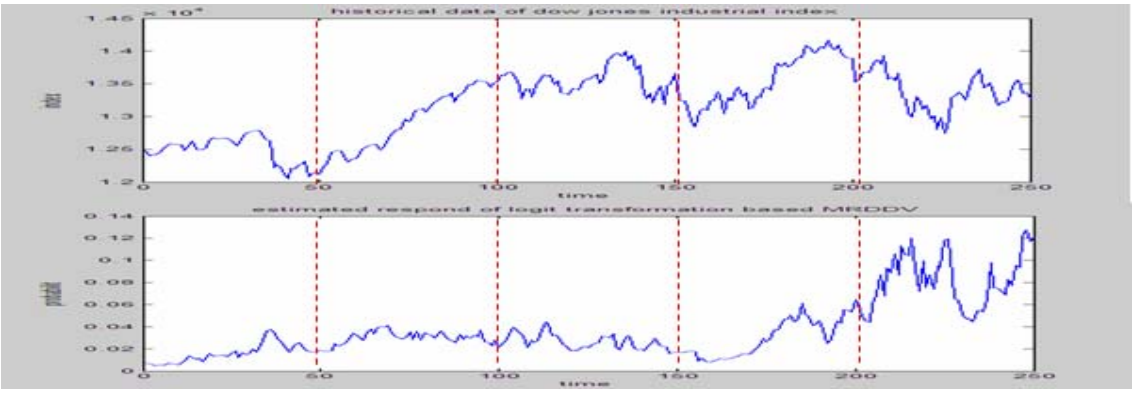


Figure 4.12: Comparison between history of DJI (top) and Logit MRDDV – DJI*DJC

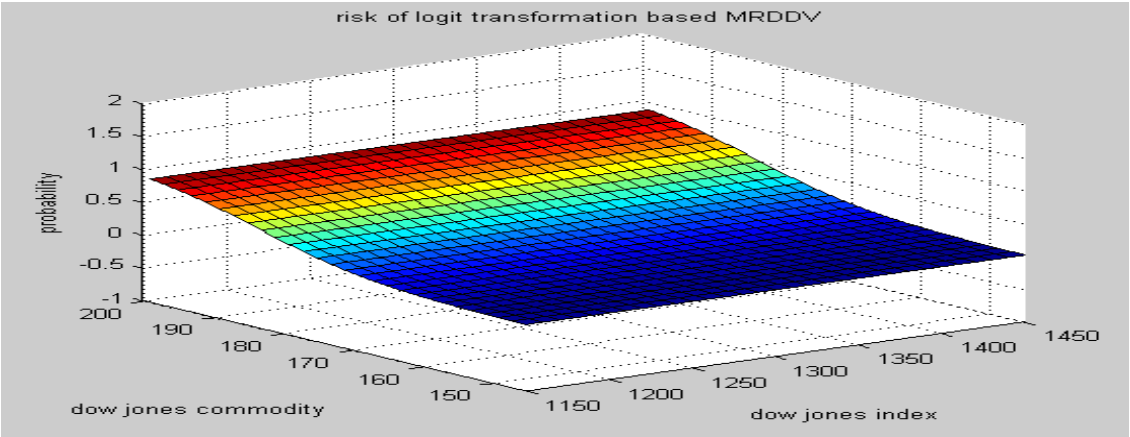


Figure 4.13: Density of Logit Transformation based MRDDV – DJI*DJC

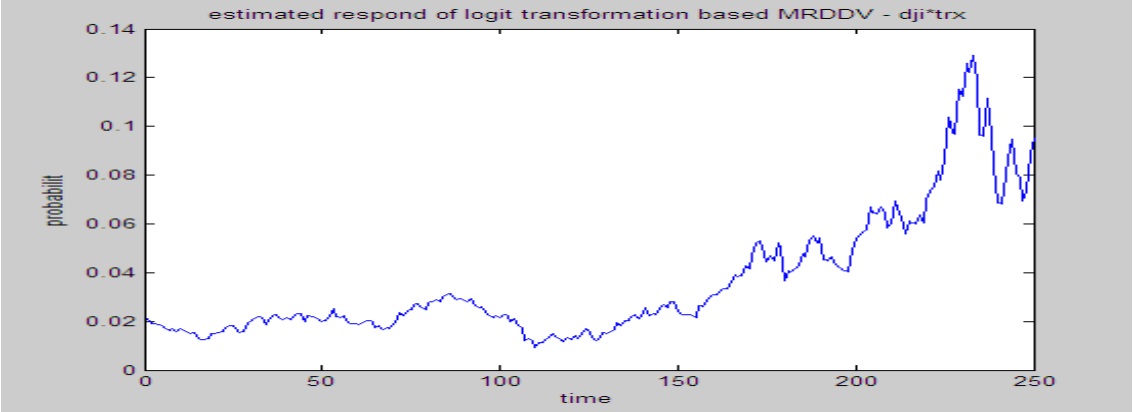


Figure 4.14: Estimated response of Logit Transformation-based MRDDV – DJI*TRX

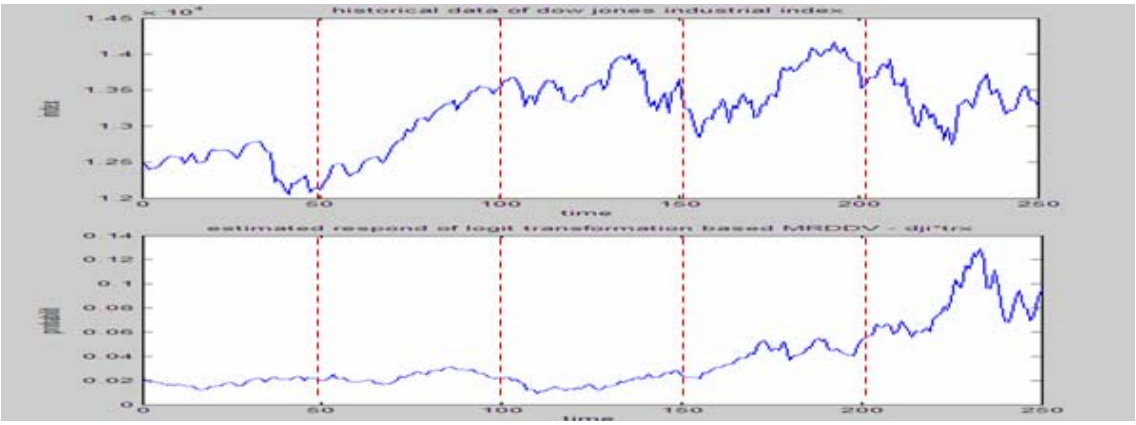


Figure 4.15: Comparison between history of DJI (top) and Logit MRDDV – DJI*TRX

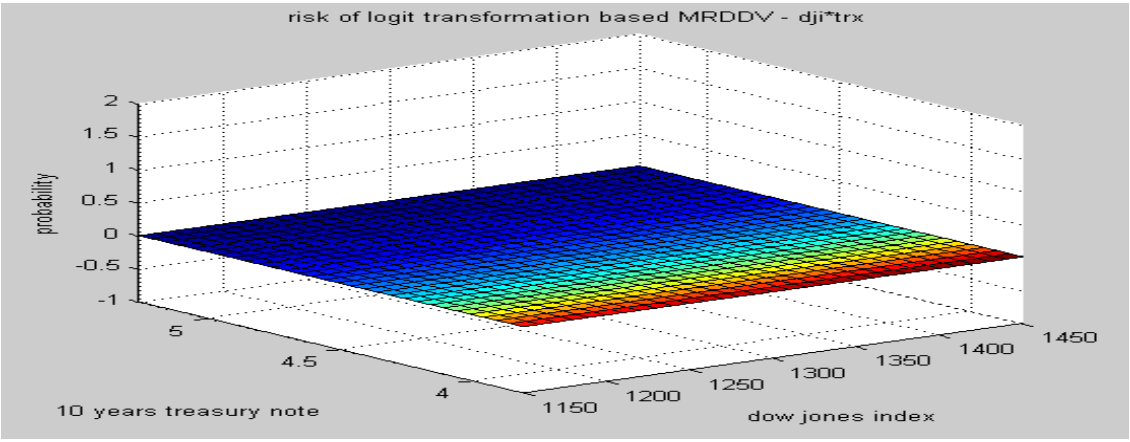


Figure 4.16: Density of Logit Transformation based MRDDV – DJI*TRX

Figure 4.10 shows the historical feature of Dow Jones Industrial Index (DJI), and Figure 4.11 - Figure 4.13 show the simulation result for the DJI and Dow Jones AIG Commodity Index (DJC). Figure 4.11 represents the estimated value of $E(y)$, that is the estimated probability computed by using the Logit Transformation-based MRDDV model, i.e., $\hat{P}(y=1)$, and this estimated value runs between 0 and 1. Figure 4.12 demonstrates a comparison between historical Dow Jones data and their estimated probability by using the Logit Transformation-based MRDDV for the DJI and DJC data, and Figure 4.13 shows the probability density function of them. In the same manner, Figure 4.14 - 4.16 demonstrate the results for the DJI and 10 Treasury Note (TRX) data.

In Figure 4.11 and 4.14, the estimated response lines drawn by the Logit Transformation-based MRDDV model represent the risks as estimated in probability such that the high probability indicates stable or relatively low chance to turn into a risky state. Notice that, as a demonstration of effectiveness of the proposed model, in Figure 4.12, at the time period around 230, the risk probability computed by the model shows a high-spike and its corresponding actual index value drastically dropped. Notice another demonstration as shown in Figure 4.15, at the time period 60 through 150, the risk probability shows a stable and low value-trend and its actual index values are moderately increasing.

Figure 4.13 and 4.16 show that the density of Logit Transformation-based MRDDV model is positive in the given domain as an evidence of correctness of the model.

CHAPTER V

CONCLUSION

This dissertation has presented a statistically-based yet probabilistically-concluded and computationally-implemented approach to modeling and evaluation of likelihood for events of interest to occur with a focus on risky events. The risky events of interest in this study are the ones with a turbulent nature in the distribution of values of data, which can be commonly found in the events in the fields such as financial market, homeland security, or safety/mission critical systems, to mention a few. In such events, it is critical to make a timely, practical and accurate forecast for the likelihood of the events of interest to occur. Two of the methods proposed in this study, i.e., Multiple Regression with a Scaled Dependent Variable (MRSDV) and Multiple Regression with a Dependent Dummy Variable (MRDDV) are multiple regression method-based, in which a quantitative and a qualitative variables are employed to represent inputs and along with an output variable to represent the consequence of the inputs on the observation through the regression process. What distinguishes MRSDV and MRDDV is the function that determines the values for the dependent variable in their models, referred to as criterion function. Using a criterion function, the dependent variable in MRSDV may result in an infinite positive range,

while that in MRDDV may result in a binary output such as 0 or 1, respectively. Another approach proposed in this dissertation is based on a new criterion function applied into the multiple regression process, referred to as Multiple Regression with an Adjusted Dependent Variable (MRADV). In MRADV, the adjustment on the dependent variable is made by determining the value of the dependent variable based on the absolute values of the forward difference of adjacent data values in the quantitative random variable in order to improve the model goodness-of-fit. Furthermore, based on the multiple regression model in the proposed MRADV, a probabilistic-based model, referred to as P_{MRADV}, has been proposed in order to derive the probability of risk (or an event of interest) to occur without relying on traditional way of assuming or establishing probability density function of the random variables in the model, thereby guiding the users to a more practical and realistic evaluation of the likelihood of an event of interest to occur. Lastly, in this dissertation, a method has been presented that can facilitate the extension of the Multiple Regression with Dependent Dummy Variable (MRDDV) Model to provide a way of estimating the likelihood of any event of concern or interest by probability. MRDDV Model employs a dependent dummy variable as an observation in its regression model with respect to the quantitative independent variable and the qualitative independent variables as primary inputs for estimation. The purpose of the dependent dummy variable in MRDDV is to provide an effective way of representing the quantitative measure of the status of the event of concern with respect to a certain criterion function, such as a binary measurement (e.g., 0 or 1) or forward differences of dependent variable values, to mention a few. Therefore, MRDDV can facilitate the process of identifying the

quantitative relationship among the random variables in the model by using the regression-based estimation. However, MRDDV lacks the ability to readily provide information on how likely an event of concern is to occur, which could be best manipulated by employing probability-based estimation. In this context, a method, namely Logit Transformation, has been employed to facilitate the probabilistic manipulation of MRDDV. By using the Logit Transformation method, the estimated dependent dummy variable can be transformed from a non-probabilistic domain (e.g., the estimated value could be in the range beyond 0 or 1) into a probabilistic one so the expected value of the dependent dummy variable can be evaluated as a probabilistic measure. The results from the Logit Transformation-based MRDDV have been extensively compared with the history of actual financial data in order to demonstrate the efficiency and effectiveness of the proposed approaches.

REFERENCES

[1] N.-J.Park, N.Park, and K.M.George, Ad-hoc Risk Management System (ARMS) using Multiple Regression with Scaled Dummy Variable (MRSDV) model: *Proc. 4th Mathmod,* Vienna, 2003, 246.

[2] N.-J.Park, N.Park, and K.M.George, "Multiple Regression with Dependent Dummy Variable (MRDDV) Model as an Ad-hoc Risk Management System (ARMS)", Proceedings of the Second IASTED International Conference on Communication Systems and Networks, Benalmadena, Spain, Sept. 8-10, 2003.

[3] N.-J. Park, K.M. George, N. Park, "Multiple Regression with Scaled Dummy Variable as a Model for Extreme Value", Society of Risk Analysis'04, Dec. 5-8, 2004, Palm Springs, CA.

[4] N.-J. Park, Nohpill Park, and K.M. George, "A probabilistic Risk Estimation with MRDDV Model using Logit Transformation", 2008 Modeling, Simulation and Visualization, Jul. 14 – 17, 2008, Las Vegas, NV.

[5] Steel, Robert G.D., Torrie, James H., Dickey, David A., Principles and Procedures of Statistics, McGraw Hill College Div. 1996.

[6] Cochran, W.G., Comparison of methods for determining stratum boundaries, Bull. Int. stat. Inst., 38, 1961, 345-358.

[7] R. V. Hogg and A. T. Craig, “Introduction to Mathematical Statistics”, Prentice Hall Inc, 5th edition, 1995

[8] Kotler, Philip, and Gary Armstrong, Principles of Marketing, 8th ed., Prentice-Hall, International Edition, 1999.

[9] S. M. Ross, “Introduction to probability models”, Academic press, 2nd edition, 1985.

[10] R.B. Nelsen, “An Introduction to Copulas”, Springer, 1999.

[11] Horvitz, D.G., Greenberg, B.G., and Abernathy, J.R., Recent developments in randomised Response designs. A survey of statistical design and linear models, American Elsevier Publishing Co., New York, 1975, 271-285.

[12] B.L.Bowerman and R.T.O;Connell, “*Forecasting and Time Series*”, 3rd edition, Duxbury, 1993.

[13] E.L. Lehmann, “Some concepts of Dependence”, The annals of Mathematical Statistics, vol. 37, no 5, pp. 1137 –1153, 1966.

[14] S. Kotz and J.P. Seeger, “A new approach to dependence in multivariate distributions”, The Netherlands: Kluwer, pp. 113-127, 1991

[15] D. Long and R. Krzysztofowicz, “A family of Bivariate Densities Constructed from Marginals”, Journal of the American Statistical Association, Vol. 90, No. 430, pp. 739-746, 1995.

[16] R.T. Clemen and T. Reilly, “Correlations and Copulas for Decision and Risk Analysis”, Management Science, Vol 45, Issue 2, pp208-224, 1999.

[17] E. J. Gumbel, Bivariate Logistic Distributions, *Journal of the American Statistical Association*, vol. 56, no. 294, pp. 335-349, 1961.

[18] J. J. Quesada-Molina, J.A. Rodriguez-Lallena, M. Ubeda-Flores, What are copulas, Monografias del Semon. Matem. Garcia de Galdeano, 27: 499-506, 2003.

[19] M. E. Johnson, A. Tenenbein, A Bivariate Distribution Family with Specified Marginals, *Journal of the American Statistical Association*, vol. 76, no. 373, pp. 198-201, 1981.

[20] R.L. Plackett, A Class of Bivariate Distributions, *Journal of the American Statistical Association*, vol. 60, no. 310, pp. 516-522, 1965.

[21] K.Denecker, S.V.Assche, J.Crombez, R.V.Vennet, and I.Lemahieu, Value-at-risk prediction using context modelling, *The European Physical Journal B, vol. 20*, 2001, 481-492.

[22] M.Radulescu, S.Radulescu, and C.Z.Radulescu, Mutiperiod portpolio selection models with transaction costs and initial holdings: *Proc 4th Mathmod* Vienna, 2003, 244.

[23] N.Gilardi, T.Melluish, and M.Maignan, Confidence evaluation for risk prediction, *Conference of the International Association for Mathematical Geology*, 2001.

[24] K.M.Thompson, R.F.Rabouw, and R.M.Cooke, The risk of grounding fatalities from unintentional airplane crashs, *Risk Analysis, Vol. 21*, No. 6, 2001.

[25] B. Dodson, "Weibull analysis", Milwaukee, Wis., ASQ Quality Press, 1994.

[26] http://www.sra.org/

[27] http://www.itl.nist.gov

[28] http://en.wikipedia.org/wiki/Main_Page

[29] http://finance.yahoo.com

[30] http://finance.yahoo.com/q/hp?s=%5EKS11, creation date: Note available, date accessed: July 12, 2006.

[31] http://finance.yahoo.com/q/hp?s=%5EKS11&a=00&b=1&c=1999&d=11&e=31&f=2005&g=m, creation date: Not Available, date accessed: February 2, 2006.

[32] http:// **www.sas.com**

APPENDICES

Theorem A.1: Parameter estimation in MRDDV model

MRDDV model is defined as $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i$, $y_i = 0,1$

where, i : number of data points and $i \geq 2$
y_i : dependent dummy variable
x_{1i} : quantitative independent variable
x_{2i} : qualitative independent variable
β_i : parameter to be estimated
e_i : error term

then, estimated parameter is $\underline{b} = (X'X)^{-1} X' \underline{y}$

where $\underline{b}$: vector of estimated values of β
X : matrix of observations of independent variable
X' : transpose of matrix X
$(X'X)^{-1}$: inverse matrix of $(X'X)$
$\underline{y}$: matrix of observation on dependent variable

Proof

i) To estimate the parameter β_i by Least Square Method:

$$S = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i}\right)^2$$

ii) S can be represented by matrix as follows:

$$\Rightarrow \quad \underline{e}' \underline{e} = \left(\underline{y} - X\underline{\beta}\right)' \left(\underline{y} - X\underline{\beta}\right) = \underline{y}\underline{y} - 2\underline{\beta}' X' \underline{y} + \underline{\beta}' X'X \underline{\beta}$$

where, $\underline{e}'$ is a transpose matrix of $\underline{e}$ and

$$X = \begin{vmatrix} 1 & x_{11} & x_{21} \\ 1 & x_{12} & x_{22} \\ : & : & : \\ 1 & x_{1n} & x_{2n} \end{vmatrix} \qquad \underline{y} = \begin{vmatrix} 0 \\ 0 \\ : \\ 1 \end{vmatrix} \qquad \underline{e} = \begin{vmatrix} e_1 \\ e_2 \\ : \\ e_n \end{vmatrix} \qquad \underline{\beta} = \begin{vmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{vmatrix}$$

iii) Differentiate $\underline{e}'\underline{e}$ with respect to $\underline{\beta}$, then solve it by letting $\underline{0}$:

⇨ $\dfrac{\partial \underline{e}'\underline{e}}{\partial \underline{\beta}} = -2X'\underline{y} + 2X'X\underline{\beta} = \underline{0}$

⇨ $X'X\underline{b} = X'\underline{y}$ where $\underline{b}$ is an estimator of $\underline{\beta}$

⇨ $\underline{b} = (X'X)^{-1}X'\underline{y}$

Theorem A.2: Property of normal random variables x_1 and x_2

If $\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_{x_1} \\ \mu_{x_2} \end{pmatrix}, \begin{pmatrix} \sigma_{x_1}^2 & \rho_{x_1x_2}\sigma_{x_1}\sigma_{x_2} \\ sym & \sigma_{x_2}^2 \end{pmatrix} \right)$

then, $b_0 + b_1x_1 + b_2x_2 \sim N\left(b_0 + b_1\mu_{x_1} + b_2\mu_{x_2},\ b_1^2\sigma_{x_1}^2 + b_2^2\sigma_{x_2}^2 + 2b_1b_2\rho_{x_1x_2}\sigma_{x_1}\sigma_{x_2} \right)$

<u>Proof</u>

i) $E(x_1) = \mu_{x_1}$, $E(x_2) = \mu_{x_2}$

➔ $E(b_0 + b_1x_1 + b_2x_2) = b_0 + b_1E(x_1) + b_2E(x_2) = b_0 + b_1\mu_{x_1} + b_2\mu_{x_2}$

ii) $V(x) = E\left[(x - E(x))^2\right] = E\left[x^2 - 2xE(x) + (E(x))^2\right]$

$= E(x^2) - 2E(x)E(x) + (E(x))^2 = E(x^2) - 2(E(x))^2 + (E(x))^2$

$= E(x^2) - (E(x))^2$

Note: $\text{cov}(x_1, x_2) = E(x_1x_2) - E(x_1)E(x_2)$, $\rho_{x_1x_2} = corr(x_1, x_2) = \dfrac{\text{cov}(x_1, x_2)}{\sigma_{x_1}\sigma_{x_2}}$

iii) $V(x_1) = \sigma_{x_1}$, $V(x_2) = \sigma_{x_2}$

➔ $V(b_0 + b_1x_1 + b_2x_2) = E\left[(b_0 + b_1x_1 + b_2x_2)^2\right] - \left(E\left[b_0 + b_1x_1 + b_2x_2\right]\right)^2$

$= b_0^2 + 2b_0b_1E(x_1) + 2b_0b_2E(x_2) + 2b_1b_2E(x_1x_2) + b_1^2E(x_1^2) + b_2^2E(x_2^2)$

$- [b_0^2 + 2b_0b_1E(x_1) + 2b_0b_2E(x_2) + 2b_1b_2E(x_1)E(x_2) + b_1^2E(x_1^2) + b_2^2E(x_2^2)]$

$= b_1^2\left[E(x_1^2) - (E(x_1))^2\right] + b_2^2\left[E(x_2^2) - (E(x_2))^2\right] + 2b_1b_2\left[E(x_1x_2) - E(x_1)E(x_2)\right]$

$= b_1^2\sigma_{x_1}^2 + b_2^2\sigma_{x_2}^2 + 2b_1b_2\rho_{x_1x_2}\sigma_{x_1}\sigma_{x_2}$

Theorem A.3: In simple regression model $y = \beta_0 + \beta_1 x + e$,

$$\hat{y} - \bar{y} = b_1\left(x - \bar{x}\right)$$

where, y : dependent variable
x : independent variable
β_i : parameter to be estimated
e : error term
$\hat{y}$: estimator of y
$\bar{y}$: mean of y
b_1 : estimator of β_1
$\bar{x}$: mean of x

Proof

i) $y = \beta_0 + \beta_1 x + e$ (transformation)

$= \beta_0 + \beta_1 x + e + (\beta_1 \bar{x} - \beta_1 \bar{x})$ (add and subtract same term $\beta_1 \bar{x}$)

$= (\beta_0 + \beta_1 \bar{x}) + \beta_1 (x - \bar{x}) + e$ (substitute $\beta_0^* = \beta_0 + \beta_1 \bar{x}$)

$= \beta_0^* + \beta_1 (x - \bar{x}) + e$

ii) $\hat{y} = b_0^* + b_1 (x - \bar{x})$ (estimation i))

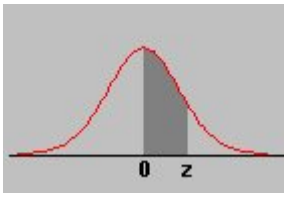
$= \bar{y} + b_1 (x - \bar{x})$ ($\because \ \beta_0^* = \beta_0 + \beta_1 \bar{x} = \bar{y}$)

iii) $\hat{y} - \bar{y} = b_1\left(x - \bar{x}\right)$

Table A.1: Probability table of standard normal

$$f(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2} z^2\right\} dz$$

Area between 0 and z



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857

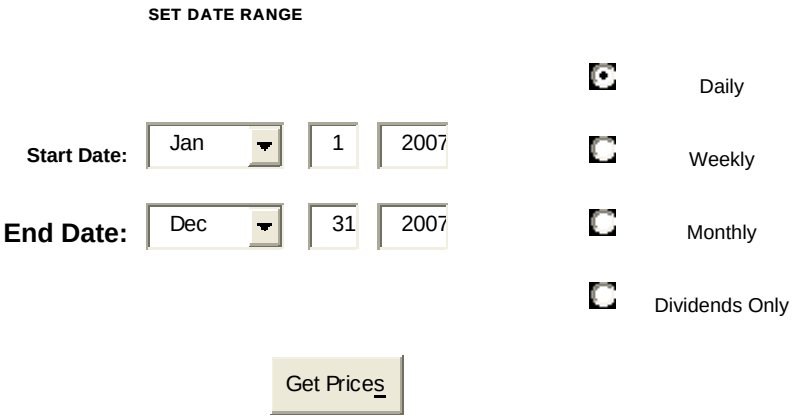
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990

Table A.2: Financial Sector Data

i) Dow Jones Industrial Average (DJI) is one of several stock market indices, created by nineteenth-century Wall Street Journal editor and Dow Jones & Company co-founder Charles Dow. It is an index that shows how certain stocks have traded.

ii) Dow Jones-AIG Commodity Index (DJC) is designed to be a highly liquid and diversified benchmark for the commodity futures market. The Index is composed of futures contracts on 19 physical commodities and was launched on July 14th, 1998.

iii) 10-Year Treasury Note (TNX) is a debt obligation issued by the U.S. Treasury that has a term of more than one year, but not more than 10 years.



Close price adjusted for dividends and splits

Date	DJI	DJC	TNX
31-Dec-07	13,264.82	184.96	4.03
28-Dec-07	13,365.87	184.77	4.10
27-Dec-07	13,359.61	185.57	4.20
26-Dec-07	13,551.69	185.38	4.28
24-Dec-07	13,550.04	182.97	4.21
21-Dec-07	13,450.65	183.03	4.17

20-Dec-07	13,245.64	180.80	4.03
19-Dec-07	13,207.27	181.37	4.07
18-Dec-07	13,232.47	179.72	4.12
17-Dec-07	13,167.20	180.10	4.19
14-Dec-07	13,339.85	180.65	4.23
13-Dec-07	13,517.96	180.97	4.17
12-Dec-07	13,473.90	183.29	4.08
11-Dec-07	13,432.77	179.55	3.99
10-Dec-07	13,727.03	178.22	4.15
7-Dec-07	13,625.58	178.34	4.12
6-Dec-07	13,619.89	178.10	4.00
5-Dec-07	13,444.96	176.22	3.91
4-Dec-07	13,248.73	176.56	3.89
3-Dec-07	13,314.57	176.52	3.89
30-Nov-07	13,371.72	177.25	3.97
29-Nov-07	13,311.73	179.20	3.94
28-Nov-07	13,289.45	179.38	4.03
27-Nov-07	12,958.44	181.27	3.94
26-Nov-07	12,743.44	184.02	3.85
23-Nov-07	12,980.88	184.24	4.01
21-Nov-07	12,799.04	181.85	4.02
20-Nov-07	13,010.14	182.35	4.05
19-Nov-07	12,958.44	179.98	4.08
16-Nov-07	13,176.79	181.76	4.15
15-Nov-07	13,110.05	180.40	4.16
14-Nov-07	13,231.01	183.10	4.27
13-Nov-07	13,307.09	180.22	4.26
12-Nov-07	12,987.55	182.01	4.21
9-Nov-07	13,042.74	184.53	4.22
8-Nov-07	13,266.29	183.56	4.27

7-Nov-07	13,300.02	183.67	4.33
6-Nov-07	13,660.94	184.92	4.36
5-Nov-07	13,543.40	182.27	4.32
2-Nov-07	13,595.10	184.19	4.29
1-Nov-07	13,567.87	182.53	4.36
31-Oct-07	13,930.01	183.52	4.47
30-Oct-07	13,792.47	180.43	4.38
29-Oct-07	13,870.26	182.79	4.38
26-Oct-07	13,806.70	180.60	4.39
25-Oct-07	13,671.92	179.22	4.35
24-Oct-07	13,675.25	176.48	4.33
23-Oct-07	13,676.23	176.61	4.41
22-Oct-07	13,566.97	177.34	4.39
19-Oct-07	13,522.02	179.05	4.40
18-Oct-07	13,888.96	180.13	4.50
17-Oct-07	13,892.54	178.51	4.55
16-Oct-07	13,912.94	178.78	4.66
15-Oct-07	13,984.80	178.97	4.67
12-Oct-07	14,093.08	176.94	4.69
11-Oct-07	14,015.12	177.21	4.66
10-Oct-07	14,078.69	175.43	4.65
9-Oct-07	14,164.53	173.33	4.65
8-Oct-07	14,043.73	172.12	4.64
5-Oct-07	14,066.01	175.73	4.64
4-Oct-07	13,974.31	177.22	4.52
3-Oct-07	13,968.05	176.60	4.54
2-Oct-07	14,047.31	176.49	4.53
1-Oct-07	14,087.55	178.08	4.56
28-Sep-07	13,895.63	178.25	4.58
27-Sep-07	13,912.94	179.71	4.57

26-Sep-07	13,878.15	177.43	4.62
25-Sep-07	13,778.65	176.71	4.61
24-Sep-07	13,759.06	177.84	4.62
21-Sep-07	13,820.19	177.21	4.63
20-Sep-07	13,766.70	177.18	4.67
19-Sep-07	13,815.56	175.80	4.52
18-Sep-07	13,739.39	175.11	4.48
17-Sep-07	13,403.42	174.54	4.47
14-Sep-07	13,442.52	172.32	4.46
13-Sep-07	13,424.88	171.74	4.48
12-Sep-07	13,291.65	172.48	4.41
11-Sep-07	13,308.39	170.24	4.36
10-Sep-07	13,127.85	168.60	4.32
7-Sep-07	13,113.38	167.12	4.37
6-Sep-07	13,363.35	166.85	4.50
5-Sep-07	13,305.47	166.91	4.47
4-Sep-07	13,448.86	167.04	4.56
31-Aug-07	13,357.74	165.57	4.54
30-Aug-07	13,238.73	165.10	4.50
29-Aug-07	13,289.29	164.43	4.55
28-Aug-07	13,041.85	163.54	4.53
27-Aug-07	13,322.13	163.99	4.60
24-Aug-07	13,378.87	164.41	4.63
23-Aug-07	13,235.88	163.14	4.62
22-Aug-07	13,236.13	161.95	4.62
21-Aug-07	13,090.86	161.06	4.59
20-Aug-07	13,121.35	161.70	4.63
17-Aug-07	13,079.08	164.21	4.67
16-Aug-07	12,845.78	161.68	4.60
15-Aug-07	12,861.47	167.33	4.71

14-Aug-07	13,028.92	167.80	4.73
13-Aug-07	13,236.53	168.14	4.78
10-Aug-07	13,239.54	167.60	4.78
9-Aug-07	13,270.68	167.52	4.79
8-Aug-07	13,657.86	168.27	4.86
7-Aug-07	13,504.30	167.84	4.74
6-Aug-07	13,468.78	167.52	4.73
3-Aug-07	13,181.91	169.60	4.70
2-Aug-07	13,463.33	170.39	4.75
1-Aug-07	13,362.37	170.82	4.76
31-Jul-07	13,211.99	172.45	4.77
30-Jul-07	13,358.31	171.46	4.80
27-Jul-07	13,265.47	170.30	4.79
26-Jul-07	13,473.57	168.74	4.78
25-Jul-07	13,785.79	169.58	4.90
24-Jul-07	13,716.95	169.46	4.94
23-Jul-07	13,943.42	170.20	4.96
20-Jul-07	13,851.08	173.48	4.96
19-Jul-07	14,000.41	174.26	5.03
18-Jul-07	13,918.22	172.81	5.01
17-Jul-07	13,971.55	169.80	5.08
16-Jul-07	13,950.98	171.10	5.04
13-Jul-07	13,907.25	174.54	5.11
12-Jul-07	13,861.73	173.51	5.12
11-Jul-07	13,577.87	173.29	5.08
10-Jul-07	13,501.70	173.30	5.04
9-Jul-07	13,649.97	171.74	5.16
6-Jul-07	13,611.68	171.63	5.20
5-Jul-07	13,565.84	170.84	5.14
3-Jul-07	13,577.30	170.10	5.05

2-Jul-07	13,535.43	170.48	5.00
29-Jun-07	13,408.62	169.67	5.03
28-Jun-07	13,422.28	168.61	5.12
27-Jun-07	13,427.73	168.95	5.07
26-Jun-07	13,337.66	168.52	5.10
25-Jun-07	13,352.05	170.25	5.08
22-Jun-07	13,360.26	170.89	5.14
21-Jun-07	13,545.84	172.40	5.16
20-Jun-07	13,489.42	173.57	5.12
19-Jun-07	13,635.42	173.57	5.09
18-Jun-07	13,612.98	176.10	5.14
15-Jun-07	13,639.48	176.48	5.17
14-Jun-07	13,553.73	174.80	5.22
13-Jun-07	13,482.35	172.17	5.20
12-Jun-07	13,295.01	171.29	5.25
11-Jun-07	13,424.96	172.27	5.14
8-Jun-07	13,424.39	169.82	5.12
7-Jun-07	13,266.73	173.07	5.10
6-Jun-07	13,465.67	173.74	4.97
5-Jun-07	13,595.46	174.81	4.98
4-Jun-07	13,676.32	175.56	4.93
1-Jun-07	13,668.11	174.03	4.96
31-May-07	13,627.64	172.72	4.89
30-May-07	13,633.08	171.44	4.88
29-May-07	13,521.34	170.08	4.88
25-May-07	13,507.28	172.37	4.86
24-May-07	13,441.13	170.47	4.86
23-May-07	13,525.65	171.58	4.86
22-May-07	13,539.95	171.88	4.83
21-May-07	13,542.88	174.82	4.79

18-May-07	13,556.53	173.27	4.80
17-May-07	13,476.72	173.47	4.76
16-May-07	13,487.53	172.81	4.71
15-May-07	13,383.84	173.23	4.71
14-May-07	13,346.78	172.29	4.69
11-May-07	13,326.22	173.44	4.67
10-May-07	13,215.13	171.10	4.65
9-May-07	13,362.87	171.77	4.67
8-May-07	13,309.07	172.01	4.63
7-May-07	13,312.97	172.85	4.64
4-May-07	13,264.62	174.17	4.64
3-May-07	13,241.38	173.75	4.67
2-May-07	13,211.88	172.41	4.65
1-May-07	13,136.14	173.10	4.64
30-Apr-07	13,062.91	173.21	4.63
27-Apr-07	13,120.94	173.56	4.70
26-Apr-07	13,105.50	171.72	4.68
25-Apr-07	13,089.89	173.85	4.65
24-Apr-07	12,953.94	171.69	4.62
23-Apr-07	12,919.40	173.49	4.65
20-Apr-07	12,961.98	172.30	4.67
19-Apr-07	12,808.63	171.61	4.67
18-Apr-07	12,803.84	171.88	4.65
17-Apr-07	12,773.04	172.12	4.69
16-Apr-07	12,720.46	172.73	4.74
13-Apr-07	12,612.13	174.35	4.76
12-Apr-07	12,552.96	173.92	4.74
11-Apr-07	12,484.62	173.86	4.74
10-Apr-07	12,573.85	173.88	4.72
9-Apr-07	12,569.14	172.42	4.74

5-Apr-07	12,560.83	173.48	4.67
4-Apr-07	12,530.05	172.90	4.65
3-Apr-07	12,510.93	170.74	4.66
2-Apr-07	12,382.30	171.47	4.64
30-Mar-07	12,354.35	171.96	4.65
29-Mar-07	12,348.75	172.33	4.63
28-Mar-07	12,300.36	170.84	4.62
27-Mar-07	12,397.29	169.62	4.61
26-Mar-07	12,469.07	169.46	4.59
23-Mar-07	12,481.01	168.98	4.61
22-Mar-07	12,461.14	169.71	4.59
21-Mar-07	12,447.52	167.18	4.52
20-Mar-07	12,288.10	166.62	4.55
19-Mar-07	12,226.17	166.57	4.57
16-Mar-07	12,110.41	166.73	4.55
15-Mar-07	12,159.68	166.57	4.54
14-Mar-07	12,133.40	166.21	4.52
13-Mar-07	12,075.96	166.12	4.49
12-Mar-07	12,318.62	167.24	4.55
9-Mar-07	12,276.32	167.84	4.59
8-Mar-07	12,260.70	169.67	4.51
7-Mar-07	12,192.45	169.20	4.50
6-Mar-07	12,207.59	167.47	4.53
5-Mar-07	12,050.41	165.93	4.52
2-Mar-07	12,114.10	168.08	4.51
1-Mar-07	12,234.34	169.76	4.56
28-Feb-07	12,268.63	171.01	4.55
27-Feb-07	12,216.24	171.98	4.51
26-Feb-07	12,632.26	173.50	4.63
23-Feb-07	12,647.48	173.39	4.68

22-Feb-07	12,686.02	172.19	4.73
21-Feb-07	12,738.41	169.84	4.69
20-Feb-07	12,786.64	167.12	4.68
16-Feb-07	12,767.57	168.27	4.69
15-Feb-07	12,765.01	166.75	4.71
14-Feb-07	12,741.86	165.69	4.73
13-Feb-07	12,654.85	166.69	4.81
12-Feb-07	12,552.55	163.86	4.80
9-Feb-07	12,580.83	167.39	4.78
8-Feb-07	12,637.63	166.36	4.73
7-Feb-07	12,666.87	164.50	4.74
6-Feb-07	12,666.31	165.48	4.76
5-Feb-07	12,661.74	165.36	4.81
2-Feb-07	12,653.49	165.11	4.83
1-Feb-07	12,673.68	164.82	4.84
31-Jan-07	12,621.69	166.09	4.83
30-Jan-07	12,523.31	164.70	4.88
29-Jan-07	12,490.78	160.60	4.89
26-Jan-07	12,487.02	163.55	4.88
25-Jan-07	12,502.56	162.38	4.87
24-Jan-07	12,621.77	163.97	4.81
23-Jan-07	12,533.80	164.88	4.80
22-Jan-07	12,477.16	161.30	4.76
19-Jan-07	12,565.53	160.41	4.77
18-Jan-07	12,567.93	157.37	4.75
17-Jan-07	12,577.15	158.61	4.79
16-Jan-07	12,582.59	158.29	4.75
12-Jan-07	12,556.08	159.56	4.77
11-Jan-07	12,514.98	156.80	4.74
10-Jan-07	12,442.16	157.30	4.68

9-Jan-07	12,416.60	155.88	4.66
8-Jan-07	12,423.49	157.07	4.66
5-Jan-07	12,398.01	157.95	4.65
4-Jan-07	12,480.69	159.21	4.62
3-Jan-07	12,474.52	161.17	4.66

VITA

Noh-Jin Park

Candidate for the Degree of

Doctor of Philosophy

Dissertation: A STUDY ON PROBABILISTIC AND COMPUTATIONAL APPROACHES TO RISK MODELING, ANALYSIS AND FORECASTING

Major Field: Computer Science

Biographical:

Personal Data: Born in Taejeon, Korea, on December 4, 1959, the son of Inbok and Hongnai Park.

Education: Graduated from Kyeongshin High School, Seoul, Korea in February 1978. Received Bachelor of Arts degree in Applied Statistics from Yonsei University, Seoul, Korea in February 1982 and Master of Science degree in Statistics from Seoul National University, Seoul, Korea in February 1984, respectively. Completed the requirements for the Doctor of Philosophy degree at Oklahoma State University in July 2009.

Experience: Employed as an associate director/ statistician; A. C. Nielsen – Korea Branch, Seoul, Korea, 1985 to 1991. Employed as a graduate research/ teaching assistant; Department of Computer Science, Oklahoma State University, 2002 to 2008. Employed as a Visiting Assistant Professor; Department of Computer Science, Oklahoma City University, 2008 to 2009.

Name: Noh-Jin Park Date of Degree: July, 2009

Institution: Oklahoma State University Location: Stillwater, Oklahoma

Title of Study: A STUDY ON PROBABILISTIC AND COMPUTATIONAL APPROACHES TO RISK MODELING, ANALYSIS AND FORECASTING

Pages in Study: 100 Candidate for the Degree of Doctor of Philosophy

Major Field: Computer Science

Scope and Method of Study: A statistically-based yet probabilistically-concluded and computationally-implemented approach to modeling and evaluation of likelihood for events of interest to occur with a focus on risky events.

Findings and Conclusions: This study introduces a method that can facilitate the extension of the Multiple Regression with Dependent Dummy Variable (MRDDV) Model to provide a way of estimating the likelihood of any event of concern by probability. MRDDV employs a dependent dummy variable in its regression model as primary inputs for estimation. However, MRDDV is not proper to provide a probability-based estimator because it violates the definition of probability. To overcome this, a method, namely Logit Transformation, is employed to facilitate the probabilistic manipulation of MRDDV. By using Logit Transformation, the estimation of risk in MRDDV is, stably, represented in probabilistic domain (e.g., in the range beyond 0 or 1). Simulation results showed that Logit Transformation-based MRDDV Model improved the basic scheme significantly. And, a user's risk defining system is, also, introduced. The enhanced Logit Transformation-based MRDDV Model is probabilistic and robust in risk tracking.

ADVISER'S APPROVAL: Dr. K. M. George